المنظمة العربية للترجمة

بول أتيويل، دايفد ب. موناغان
بمشاركة دارين كوونغ

مدخل إلى التنقيب في بيانات العلوم الاجتماعية

ترجمة

عبد النور خراقي

سلسلة كتب منتقاة في علم الاجتماع (١)

مدخل إلى التنقيب

مدخل إلى التنقيب
في بيانات العلوم الاجتماعية

لجنة العلوم الإنسانية والاجتماعية

هدى مقنص (منسقةً)

سمية الجرّاح

رجاء مكي

صالح أبو إصبع

عمر الديوه جي

مصطفى حجازي

المنظمة العربية للترجمة

بول أتيويل، دايفد ب. موناغان
بمشاركة دارين كوونغ

مدخل إلى التنقيب في بيانات العلوم الاجتماعية

ترجمة
عبد النور خراقي

مراجعة

هيثم الناهي   عدنان عيدان

الفهرسة أثناء النشر - إعداد المنظمة العربية للترجمة
أتيويل، بول
مدخل إلى التنقيب في بيانات العلوم الاجتماعية/ دايفد ب. موناغان بمشاركة دارين كوونغ؛ ترجمة عبد النور خراقي؛ مراجعة هيثم الناهي وعدنان عيدان.
392 ص. - (علوم إنسانية واجتماعية)
بيبليوغرافيا: 377-386.
يشتمل على فهرس
ISBN 978-614-434-105-6
1. البيانات. 2. الاجتماع، علم. أ. العنوان. ب. موناغان، دايفد ب. (مؤلف). ج. كوونغ، دارين (مؤلف). د. خراقي، عبد النور (مترجم). هـ. الناهي، هيثم (مراجع). و. عيدان، عدنان (مراجع) ز. السلسلة.
300

«الآراء الواردة في هذا الكتاب لا تعبّر بالضرورة
عن اتجاهات تتبناها المنظمة العربية للترجمة»

Attewell, Paul and David B. Monaghan with Darren Kwong
Data Mining for the Social Sciences: An Introduction

المنظمة العربية للترجمة

بناية «شاتيلا وقهوجي»، شارع ليون، ص. ب: 113 - 5996
الحمراء - بيروت 2090 1103
لبنان، هاتف: (9611)753031 - 753024
فاكس:(9611) 753032
e-mail: aotarab@gmail.com
http://www.aot.org.lb

الطبعة الأولى: بيروت، تموز/ يوليو 2020
يمكنكم شراء هذا الكتاب عبر الموقع الإلكتروني: **www.aot.org.lb**

المحتويات

الجزء الثاني: أمثلة عملية

مدخل إلى التنقيب

في بيانات العلوم الاجتماعية

مقدمة المترجم

لا شكّ في أنّ أي مشروع بحث علمي يعتمد التحليل والتمحيص، للإجابة عن أسئلة شائكة، يتوسل بطرق تحليلية تتوخى قدراً كبيراً من الدقة، بغية منح نتائجه مصداقية ومرجعية متميزتين. ولعل الاهتداء إلى استنباط الأنماط المفيدة، ذات الصلة الوثيقة بأهداف المشروع البحثي داخل بيانات ضخمة، يضيع في تفاصيلها الباحث، هو المفتاح الرئيس نحو تحقيق هذا المبتغى. ضمن هذا التصور العام، يقدم بول أتيويل ودايفد موناغان مدخلًا مفيداً في التنقيب في البيانات، الذي يشير إلى إحدى أهم الطرق الحديثة في التعامل مع معالجة البيانات، ورصد الأنماط الهامة المتصلة بغاية البحث.

إن التنقيب في البيانات، أو ما يطلق عليه أحياناً اسم استكشاف البيانات أو المعرفة، عملية من عمليات تحليل البيانات، وتلخيصها ضمن معلومات مفيدة، قد تستخدم مثلاً في زيادة الدخل، أو تخفيض التكاليف أو هما معاً. وإنّ برمجيات التنقيب في البيانات هي إحدى الوسائل التحليلية العديدة المسخرة في عملية التنقيب في البيانات؛ فهي تمكن المستخدمين من تحليل البيانات انطلاقاً من أبعاد ورؤى مختلفة، وتصنيفها، وتلخيص العلاقات المرصودة. ومن الناحية التقنية، يعد التنقيب في البيانات، عملية تحدد الارتباطات (Correlations) أو الأنماط الموجودة بين عشرات الحقول في قواعد البيانات العلائقية (Relational Databases) الضخمة.

صحيح إن التنقيب في البيانات هو اصطلاح جديد، ولكن التقنية مألوفة، ذلك بأن الشركات سبق أن استعملت حواسيب قوية في غربلة أحجام كبيرة من بيانات الماسح الضوئي للأسواق الضخمة، وتحليل تقارير بحثية عنها. ومع ذلك يبقى هذا التحليل محدوداً بالمقارنة مع ما وصلت إليه الابتكارات المستمرة في مجال المعالجة الحاسوبية، وتخزين القرص، والبرمجيات الإحصائية التي رفعت من دقة تحليل البيانات على نحو لافت للنظر. وقد تكون البيانات وقائع، أو أعداداً، أو نصوصاً يمكن أن يخضع إلى المعالجة الحاسوبية، كما أن التقدم الذي تم تحقيقه في مجال برمجيات الحاسوب، مكنت المنظمات والشركات، وغيرها، من دمج قواعد بياناتها في مستودع البيانات (Data Warehouse)، إذ تدار داخله البيانات بشكل منظم وتسترجع متى شاء المحلل ذلك. ومن بين هذه البرمجيات التحليلية، نذكر البرمجيات الإحصائية، وبرمجيات التعلم الآلي (Machine Learning)، وبرمجيات الشبكات العصبية، بحيث تسعى كلها إلى البحث في «الأصناف» (Classes)، و«التجميعات» (Clusters)، و«الترابطات» (Associations)، «والأنماط التسلسلية» (Sequential Patterns).

ولدى التنقيب في البيانات، مستويات مختلفة من التحليل كالشبكات العصبية الاصطناعية، والخوارزميات الجينية، وتفرعات القرار، وطريقة أقرب الجيران، واستقراء القاعدة، وتصور البيانات، وغيرها من المستويات والطرق التحليلية.

لقد ظل المؤلفان – من أولى كلمات الكتاب إلى نهايتها – يدافعان بحماس عن التنقيب في البيانات باعتبارها طريقة أو مقاربة بديلة عن النمذجة الإحصائية التقليدية، التي تعجز عن معالجة البيانات الضخمة، والمألوفة لدى معظم علماء الاجتماع.

وقبل أن أختم هذه المقدمة المقتضبة، لا بُدَّ من الإشارة إلى المشاكل الجمّة التي رافقتني طيلة القيام بترجمة هذا الكتاب العلمي الهام. لما عرض عليَّ كتاب *Data Mining For The Social Sciences: An Introduction* شدّ انتباهي عنوانه، وشغلت تفكيري عبارة Data Mining، بخاصة. حاولت أن أُترجمها دون اللجوء

إلى محتوى الكتاب برمته، فعجزت؛ وبعد الاطّلاع على الكتاب ومراميه، حاولت مع ذلك الاستئناس ببعض الجهود الترجمية التي تعرضت لهذه الكلمة الحبلى بالمعاني التقنية، فوجدت مَنْ ترجمها بعبارة «استنباط البيانات»، ومن ترجمها بعبارة «التنقيب عن البيانات». لم تقنعن أي من الترجمتين، ذلك بأن الأولى تهمل معنى التنقيب الذي استعاره الكاتبان لإيصال فكرتهما، والثانية تذكر كلمة التنقيب المطلوبة في إبراز ما يرومه المؤلفان، غير أن استعمالها اللغوي الذي يتبع حتماً بحرف «عن»، يوحي للقارئ بأن البيانات قيد الدرس غير موجودة أصلاً، ومن ثم، وجوب جمعها. أمام هذا القصور في فهم العبارة، وترجمتها ترجمة تلتزم بروح المعنى الذي يتوخاه الكاتبان، اقترحت عبارة التنقيب في البيانات، التي تقتضي وجود بيانات في المقام الأول، تخضع للتنقيب بغية فهم ما بها من أسرار تحليلية هيكلية.

إنَّ ترجمة النص العلمي الذي قد تترتب عنه اختراعات وبناء تصورات، خطيرة جداً، خطورة ترجمة النص الديني أحياناً، ولهذا كانت معظم قراراتي المتعلقة بانتقاء الأنسب من المقابلات العربية، صعبة للغاية؛ فالمصطلحات العلمية (الرياضية منها، والحاسوبية، والإحصائية بخاصة)، جديدة على الساحة العلمية، وتتطلب من الباحث المترجم ذكاءً استثنائياً لنحت مقابلاتها في اللغة العربية؛ لا أُخفِ القارئ أن رحلتي كلها مع هذا الكتاب المتفرد في الهدف والشكل، كانت رحلة شكّ في كُلّ كلمة مدرجة بشكل مستقل أو مضافة، سواء كانت سهلة جداً أو متباينة الصعوبة، ولهذا تراني أحياناً أقترح المقابل وأتبعه بكتابته بالحروف الإنجليزية (Transliteration).

وأخيراً أشكر المنظمة العربية للترجمة في شخص مديرها العام أ. د. هيثم الناهي، الذي منحني كُلّ هذه الثقة للتصدي لكتاب علمي من هذا العيار الثقيل. كما أشكر زوجتي التي شجعتني على ترجمة الكتاب دون تردد، ووفرت لي الأجواء المناسبة لإتمامه.

عبد النور خراقي

إهداء

إلى عائلتي، كاتي، وتيفان، ودايفد، الذين دفعني
دعمهم ومودتهم إلى كُلّ ما قمت به.

بول أتيويل

إلى زوجتي الرائعة، ميليندا على حبها،
ودعمها، وتشجيعها. وإلى والديّ على حبهما، وتوجيههما.

دايفد ب. موناغان

شكر وتقدير

إن التنقيب في البيانات - خاصة باعتباره تخصصاً يُطبَّق على بيانات العلوم الاجتماعية - هو مجال بحث، يعرف تغيراً متسارعاً. واستفاد فهمنا لهذه الطرق الجديدة بشكل هائل من تعليم الآخرين ونصحهم، خصوصاً الأستاذ روبرت ستاين، وروبرت هاراليك، وأندرو روزنبورغ. هذا، وقد ساهم العديد من الطلاب، ممن يستعملون هذه التقنيات في مشاريع الدكتوراه، بحكمتهم.

أولاً وقبل كُلّ شيء، أَعدّ دارين كوونغ العديد من الأمثلة التي وردت في هذا الكتاب، متصارعاً أحياناً أثناء مباشرة العمل مع عناد البرمجيات، كما نظم دارين أيضاً سلسلة ندوات عامة، لا تقدر بثمن عن التنقيب في البيانات في مركز كوني للدراسات العليا في نيويورك التي تطلعنا على الطرق الكمية الجديدة. وقد شاطر كُلّ من ديرك ويتيفين وأندرو والاس استبصاراتهما ومهاراتهما حول تقنيات حاسوبية مختلفة، يتقنونها. وساهمت وينغوان وتشنغ بعملها الجاد، لإتمام مهمة لا تبغي من ورائها شكراً خاصاً. وثمة طلبة متخرجون آخرون كثر، لا يمكن ذكرهم جميعاً، المسجلين في دورات التكوين بسلك الدكتوراه في مجال التنقيب في البيانات، والذين منحونا فرصة اختبار أفكارنا وشروحاتنا لهذه الطرق.

وأخيراً وليس آخراً، إننا مدينون بالشكر الجزيل لمؤسسة العلوم الوطنية، التي دعمت منحتها التي تحمل رقم DRL1243785، بحثنا وأنشطة أخرى ذات الصلة، بما في ذلك التنقيب في البيانات في العلوم الاجتماعية والسلوكية، وفي التعليم.

الجزء الأول

مفاهيم

الفصل الأول

ما المقصود بالتنقيب في البيانات؟

يطلق اسم التنقيب في البيانات (Data Mining) (DM) على مجموعة من تقنيات الحاسوب المكثف، بغية استكشاف البنية، وتحليل الأنماط في البيانات. ومن خلال استخدام تلك الأنماط، يمكن للتنقيب في البيانات أن ينتج نماذج تنبؤية، أو يصنف الأشياء، أو يحدد مجموعات أو تجميعات (Clusters) مختلفة من الحالات داخل البيانات. وقد سبق استخدام التنقيب في البيانات، وبطرق أُخرى مثل التعلم الآلي (Machine Learning)، والتحليلات التنبؤية (Predictive Analytics)، في الاتّجار بشكل واسع، وأخذ ينتشر في العلوم الاجتماعية، ومجالات بحث أخرى.

وتضم القائمة الجزئية لمناهج التنقيب في البيانات الحالية ما يلي:

- قواعد الارتباط (Association Rules)
- تقسيم تكراري (Recursive Partitioning) أو أشجار القرار (Decision Trees)، بما في ذلك التصنيف وشجرة الانحدار (Classification and Regression Trees) (CART)، ومربع كاي للكشف عن التفاعل التلقائي (Chi-Squared Automatic Interaction Detection) (CHAID)، وأشجار معززة (Boosted Trees)، وغابات، وغابات نظام تمهيدي لتشغيل الحاسوب (Bootstrap Forests).

• نماذج الشبكة العصبية المتعددة الطبقات (Multi - Layer Neural Network Models) ومناهج «التعلم العميق» (Deep Learning).

• مصنفات «بايز» (Bayes Classifiers) الساذجة، والشبكات «البايزية» (Bayesian Networks).

• المناهج التجميعية، (Clustering Methods) بما في ذلك أقرب المجاورات التراتبية خوارزمية «ك-مينز» (k-Means)، والتجميع المتعدد الخطي وغير الخطي.

• شعاع الدعم الآلي (Support Vector Machines).

• «نمذجة لينة» (Soft Modeling) أو نمذجة متغيرة المربعات الصغرى الكامنة (Partial Least Squares Latent).

يُعد التنقيب في البيانات علم حديث العهد، ولكنه ينمو نمواً فائق السرعة، إذ تظهر - في اللحظة الراهنة من حديثنا - طرق جديدة، وتعدل طرق قديمة، وتتراكم استراتيجيات ومهارات تمكّن من استخدامها. لقد أصبحت قوة التنقيب في البيانات وأهميتها تحظى باعتراف واسع النطاق، إذ في غضون السنتين الماضيتين فقط، ضخت المؤسسة الوطنية للعلوم، ملايين الدولارات للنهوض بمبادرات بحث جديدة في هذا المجال.

ويمكن تطبيق طرق التنقيب في البيانات على ميادين مختلفة جداً، مثل البيانات المرئية، أو قراءة خطّ اليد (القراءة الضوئية للحروف)، أو التعرف على الوجوه داخل صور رقمية. كما يستخدم التنقيب في البيانات في تحليل النصوص (مثل تصنيف مضمون المقالات البحثية أو وثائق أخرى)، ومن ثم ظهور عبارة التنقيب في النصوص (Text Mining). علاوة على ذلك، يمكن تطبيق تحليلات التنقيب في البيانات على التسجيلات الصوتية (Digitized Sound) للتعرف - مثلاً - على كلمات تَرِدُ في محادثات هاتفية. ولكننا سنركز في هذا الكتاب على المجال الأكثر شيوعاً: استخدام طرق التنقيب في البيانات لتحليل البيانات الكمية (Quantitative Data) أو الرقمية.

إن عمال المناجم ينقبون عن عروق المعدن الخام، ويستخرجون هذه الأجزاء النفيسة من الصخور المحيطة. وقياساً على ذلك، يسعى التنقيب في البيانات إلى التنقيب عن أنماط أو بنية في البيانات. ولكن ماذا نقصد عند قولنا إننا ننقب عن بنية داخل بيانات؟ تصور شاشة حاسوب ما، التي تعرض آلاف البيكسلات، أي نقاط الضوء أو الظلام (Pixels)، التي تعد بيانات خامة أو أولية. ولكن لو فحصت تلك البيكسلات عبر العين، وتعرفت - في داخلها - أشكال الحروف والكلمات، فإنك بصدد إيجاد بنيات في البيانات - أو لنستخدم استعارة أخرى، فإنك بصدد تحويل البيانات إلى معلومة (Information).

إنَّ مقابل شاشة الحاسوب بالنسبة إلى البيانات الرقمية، جدول ممتد (Spreadsheet) أو مصفوفة (Matrix)، بحيث تمثل كُلّ خانة متغيراً (Variable) واحداً، وكل سطر (Row) يضم بيانات بالنسبة إلى شخص أو حالة مختلفين. كما تضم كُلّ خلية داخل الجدول الممتد، قيمة محددة بالنسبة إلى شخص واحد بخصوص متغير معين.

كيف يتسنى إدراك الأنماط، أو الانتظام، أو البنية في هذا النوع من البيانات الأولية الرقمية؟ يقدم علماء الإحصاء طرقاً متنوعة للتعبير عن العلاقات القائمة بين الخانات والأسطر في جدول ما، والمصفوفة الترابطية (Correlation Matrix) هي إحدى هذه الطرق الأكثر شيوعاً. وعوضاً عن ترديد (Repeating) البيانات الأولية (Raw Data)، المؤلفة من آلاف الملاحظات، وعشرات المتغيرات يمكن أن تمثل المصفوفة الترابطية مجرد العلاقات بين كُلّ متغير، وكل متغير آخر على حدة. إنها ملخص، أي إنها تبسيط للبيانات الأولية.

القليل منا مَنْ يستطيع قراءة المصفوفة الترابطية ببساطة، أو يدرك نمطاً هادفاً فيها، باستثناء قلة قليلة. من أجل هذا، نتوسل - إجمالاً - بخطوة ثانية للبحث عن بنيات في بيانات رقمية؛ فنبتكر نموذجاً يلخص العلاقات في المصفوفة الترابطية، مثل نموذج انحدار المربعات الصغرى (Ordinary Least Squares Regression)، الذي يترجم هذه المصفوفة الترابطية إلى معادلة انحدار (Regression Equation) متناهية في الصغر، يمكننا فهمها وتفسيرها بسهولة أكثر.

ومع ذلك، يعد نموذج إحصائي لا أكثر من مجرد كونه تلخيصاً مشتقاً من بيانات أولية، بل هو أيضاً أداة للتنبؤ (Prediction)، وهي الخاصية التي تجعل من التنقيب في البيانات مفيدة، خاصة، إن البنوك تُراكم بيانات ضخمة (Huge Data) حول الزبائن، بما في ذلك تسجيلات تهمّ أولئك المتخلفين عن الإيفاء بتسديد القروض، وإذا ما تمكن محللون مصرفيون من تحويل تلك البيانات إلى نموذج يسمح على نحو دقيق، بمن سيتخلف عن أداء قرض ما، فسيكون باستطاعتهم رفض الطلبات الجديدة الأكثر مجازفة بشأن الحصول على القروض، ومن ثم تجنب الخسائر. وإذا ما تمكنت شركة أمازون كوم (Amazon.com)، من تقييم الأذواق بشأن الكتب التي تستهوي الميول الشخصية، استناداً إلى المقتنيات السابقة، مع رصد أوجه التطابق بين عملاء آخرين، ومن ثم العمل على الإغراء بعرض كتب مختارة بعناية، فتحقق هذه الشركة مزيداً من الأرباح. وإذا ما تمكن طبيب ما، من الحصول على تَفْريسة بالرنين المغناطيسي النووي (NMR Scan)، لنسيج الخلايا، والتنبؤ - انطلاقاً من تلك البيانات - بما إن كان ورم ما، خبيثاً أم حميداً، فستكون رهن إشارة الطبيب، أداة قوية.

إن عالمنا يعج بالبيانات الرقمية، ومن خلال عملية التنقيب فيها، بغية إيجاد أنماط ما - خاصة أنماط قادرة على التنبؤ بنتائج مهمة بشكل دقيق - يمكنها تقدم خدمة قيمة للغاية. فالتنبؤ الدقيق، يمكن أن ينذر بقرار، ويفضي إلى العمل على اتّخاذه. وإذا كان ذلك النسيج الخلوي خبيثاً على الأرجح، فلا بُدَّ للمرء - إذن - من برمجة عملية جراحية؛ وإذا كانت نسبة الخطر المتوقعة عالية بشأن تخلف الدائن عن أداء القروض، فلا تقرضه.

ولكن لماذا الحاجة إلى التنقيب في البيانات من أجل هذا؟ أليست هذه الطرق الإحصائية التقليدية غير قادرة على القيام بأداء الوظيفة نفسها على أتم وجه؟

لا شك في أن الطرق الإحصائية التقليدية تمنح نماذج تنبؤية، غير أنها لا تسلم من نقص كبير. من أجل ذلك، ظهرت طرق التنقيب في البيانات باعتبارها بديلاً عن الطرق التقليدية، وأحياناً بديلاً أفضل، أقل ارتهاناً بتلك المشاكل. وسنقوم لاحقاً بتعداد مزايا متعددة للتنقيب في البيانات، غير أننا نقتصر حالياً على الميزة الأكثر وضوحاً. إن التنقيب في البيانات مناسب خاصة، لتحليل مجموعات بيانات

(Datasets) كبيرة جداً ذات متغيرات و(أو) حالات عديدة، تعرف بالبيانات الضخمة (Big Data).

وأحياناً تنهار طرق الإحصاء التقليدية لدى تطبيقها على مجموعات كبيرة جداً من البيانات، ومرد ذلك، إما إلى عجزها عن معالجة مظاهر حاسوبية، وإما إلى مواجهتها عوائق أكثر جوهرية في التقدير عندما تحتوي - مثلاً - مجموعة بيانات على متغيرات تفوق الملاحظات؛ وهو مزج تعجز نماذج الانحدار التقليدية عن معالجته، ولكن تتوفق فيه طرق عديدة من التنقيب في البيانات.

لا يقتصر التنقيب في البيانات على التغلب على بعض النقص الذي تعاني منها طرق الإحصاء التقليدية، بل تساعد أيضاً على تجاوز بعض النقص البشري. وقد يغفل باحث ما سمات مهمة من البيانات، وهو يواجه مجموعة بيانات مكونة من مئات المتغيرات وآلاف مؤلفة من الحالات، بالنظر إلى قلة الوقت والانتباه. على سبيل المثال، من السهل نسبياً، فحص ست متغيرات للبث في تحويل أي منها، وجعلها أكثر تطابقاً مع منحنى جرسي (Bell Curve)، أو توزيع طبيعي (Normal Distribution). ومع ذلك، سيصاب محلل بشري ما بالارتباك بشكل سريع لدى محاولته تطبيق الأمر نفسه على مئات المتغيرات. وعلى النحو ذاته، قد يرغب باحث ما في فحص تفاعلات إحصائية بين متنبئين في مجموعة بيانات معينة، ولكن ما الذي سيحدث لمّا يكون ذاك الشخص ملزماً بالأخذ بعين الاعتبار تفاعلات بين عشرات المتنبئين؟ إن عدد التركيبات المحتملة تنمو بشكل كبير جداً، إلى درجة أن أي محلل بشري يجد نفسه في وضع لا يحسد عليه.

وتعد تقنيات التنقيب في البيانات - في هذه الحالة - مفيدة، لأنها تساعد جزئياً على «أتمتة» (Automate) تحليل البيانات، من خلال تحديد المتنبئات الأكثر أهمية بين عدد كبير من المتغيرات المستقلة، أو من خلال تحويل المتغيرات آلياً، إلى توزيعات أكثر فائدة، أو عبر اكتشاف التفاعلات المعقدة بين المتغيرات، أو عبر استجلاء الأشكال غير المتجانسة السائدة في مجموعة بيانات ما. ويتخذ الباحث البشري قرارات حاسمة، ولكن طرق التنقيب في البيانات تؤثر في قدرة الحواسيب على مقارنة بدائل عديدة،

وتحديد أنماط قد يهملها المحللون من البشر بسهولة (Larose 2005; Mckinsey Global Institute 2011; Nibset, Elder, and Miner 2009).

ومحصلة ذلك أن التنقيب في البيانات كثيف جداً حسابياً، ذلك بأنه يستخدم قدرة الحاسوب للتنقيب عن البيانات بغية استخلاص أنماط معينة، والبحث عن التفاعلات «الخفية» بين المتغيرات، واختبار طرق بديلة أو مزج نماذج لتعظيم دقة تنبؤه.

أهداف هذا الكتاب

ثمة كتب عديدة عن التنقيب في البيانات؛ فبماذا يمتاز هذا الكتاب عن غيره، إذن؟ قد يفكر المرء في أدبيات حول التنقيب في البيانات، باعتبارها كعكة مكونة من عدة طبقات، حيث تتناول طبقتها السفلى التصورات والنظريات التي تشكل الدعامة الأساسية للتنقيب في البيانات. هذه أمور جوهرية، ولكنها مستعصية على الفهم. ولم يكن هدف هذا الكتاب الرئيس، تناول الأشياء تقنياً على مستوى عالٍ جداً، ولكن يمكن للمهتمين من القراء الاطّلاع على جوانب من ذلك، من خلال الرجوع إلى النسخة الإلكترونية من النص الكلاسيكي من إنتاج هاستي (Hastie)، وتيبشيراني (Tibshirani)، وفريدمان (Friedman): (*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*) (2009) وتوجد نسخة مجانية على الرابط التالي:

(www.stanford.edu/~hastie/local.ftp/Springer/OLD//ESLII_print4.pdf)

وإذا ما تحركنا تصاعدياً، فسنجد الطبقة الموالية من أدبيات التنقيب في البيانات، المتضمنة الخوارزميات (Algorithms) الحاسوبية التي تطبق تلك التصورات الرياضية على البيانات. وتتجلى القضايا الجوهرية في هذا السياق، تتجلى في تقليص الوقت المطلوب لأداء عمليات رياضية ومصفوفة (Matrix)، واختيار الاستراتيجيات الحاسوبية الناجعة، القادرة على تحليل حالة واحدة على حدة، أو القيام بعدد محدود جداً من التنقلات عبر مجموعة بيانات ضخمة. وتكون استراتيجيات الحاسوب

الناجعة، حاسمة بخاصة - وبشكل سريع - عند تحليل بيانات ضخمة، تتألف من مئات الآلاف من الملاحظات. ويمكن أن يشتغل برنامج حاسوب غير ناجع لأيام لإنجاز تحليل واحد. وهذا الكتاب لا يخوض في المستوى الخوارزمي البتة؛ والقراء المهتمون، يمكنهم الرجوع إلى كتب تان (Tan)، وشتاينباخ (Steinbach)، كومار (Kumar) (2005) وويتن (Witten)، وإيبي (Eibe)، وهول (Hall) (2011).

وفي الطبقة العليا من أدبيات التنقيب في البيانات، يجد المرء كتباً حول استخدام التنقيب في البيانات؛ إذ يحمل كثير منها نصائح للمديرين والموظفين، تمكنهم من إحداث ثورة شاملة في شركاتهم من خلال تبني التنقيب في البيانات أو «تحليل الأعمال» باعتبارها استراتيجية عمل. ومع ذلك، ليس ذلك هدفنا، بل إن هذا الكتاب يقدم مدخلاً قصيراً غير تقني لأولئك الذين يهتمون باستخدامها في تحليل بيانات كمية، ولا يعرفون - مع ذلك - الكثير عن هذه الطرق. إن هدفنا الرئيس هو تفسير عمل التنقيب في البيانات، وكيفية اختلافها عن أنواع مألوفة أو راسخة للغاية، من التحليل الإحصائي والنمذجة (Modeling)، والوقوف عند بعض مواطن القوة والضعف التي يتميز بها التنقيب في البيانات. ولبيان تلك الأفكار، يبدأ الكتاب بمناقشة التنقيب في البيانات بشكل عام، لا سيما ما يتعلق بمنظوره المتميز حول تحليل البيانات؛ وتنتقل المناقشة بعد ذلك، إلى تقديم الطرق الرئيسة أو الأدوات داخل التنقيب في البيانات.

ويتحاشى الكتاب - في مجمله - الرياضيات، ولكن يفترض معرفة أساسية بالإحصائيات التقليدية، ويفرض - على الأقل - الإلمام بقدر ضئيل بالانحدار المتعدد[1] (Multiple Regression)، والانحدار اللوجيستي (Logistic Regression). ويقدم القسم الثاني من هذا الكتاب، أمثلة عن تحليلات البيانات بالنسبة إلى كُلّ تطبيق على حدة أو أداة من أدوات التنقيب في البيانات، كما يطلع الكتاب القارئ على تأويل مخرجات البرمجيات (Software Output)، ويناقش كُلّ مثال من الأمثلة التي علمتنا. ويضم هذا الكتاب «حيلاً» عديدة، يستخدمها محللو البيانات في تحليلاتهم، ويبرز بعض المآزق قصد تجنبها، أو يقترح طرقاً لاحتوائها.

(1) يمكن أيضاً ترجمة هذه العبارة بـ «التراجع المضاعف» (المترجم).

وبعد الانتهاء من قراءة هذا الكتاب، تكون مطالباً - على نحو عام - بفهم معنى التنقيب في البيانات، وإدراك غايات استخدامها من لدن محلل البيانات، وتكون قادراً على اختيار أدوات التنقيب في البيانات المناسبة من أجل القيام بمهام خاصة، وقادراً أيضاً على تفسير مخرجاتها. ويبقى بعد ذلك استخدام أدوات التنقيب في البيانات - بالأساس - مسألة ممارسة، ومسايرة لحقل يشهد تقدماً بوتيرة متسارعة وعلى نحو غير عادي.

برمجيات ومعدات من أجل التنقيب في البيانات

تستخدم شركات كبيرة برامج الحاسوب المكتوبة للعملاء في تطبيقات (Applications) التنقيب في البيانات، ويشغلونها مستخدمين الحاسبات الكبرى (Mainframes) فائقة السرعة، أو تجميعات حاسوبية[2] (Computer Clusters) قوية. وتعد - على ما يبدو - تلك الأنواع من الحواسيب، أفضل الحواسيب البيئية المستعملة في تحليل البيانات الضخمة (Big Data)، ولكن ليست في متناول السواد الأعظم منا. ولكن، لحسن الحظ أن هناك منتوجات متعددة، تمزج أدوات متعددة للتنقيب في البيانات، في حزمة واحدة أو مجموعة برمجيات (Software Suite)، يتم تشغيلها ضمن نظام ويندوز (Windows) على حاسوب شخصي.

إن جي. أم. ب (JMP) التي تنطق «غامب برو»، وهي برمجيات إحصائية من تطوير الشركة التي تبيع برمجيات نظام التحليل الإحصائي (SAS)، يمكن للمرء تحميل نسخة تجريبية منها بالمجان. كما تقدم الشركة برامج تعليمية عبر الإنترنت، وأدوات تعليمية أخرى. إن برمجيات «غامب برو» سهلة الاستخدام نسبياً بواسطة استعمال منهجية الإشارة والنقر (Point-and-Click Approach). ومع ذلك، فهي تفتقر إلى بعض أدوات التنقيب في البيانات التحليلية الأكثر حداثة.

وتعد الحزمة الإحصائية للعلوم الاجتماعية (Statistical Package for the Social Sciences (SSPS)) التي تملكها شركة آي. بي. إم (IBM)، أقدم المنتوجات البرمجية، وأكثرها رسوخاً في تحليل البيانات، متوسلة بطرق إحصائية تقليدية مثل

---

(2) يجوز أيضاً ترجمة العبارة بـ «عناقيد حاسوبية» (المترجم).

الانحدار (Regression)، والتبويب المزدوج (Cross-Tabulation)، و«اختبار-ت» (T-Test) (أي اختبار المقارنة بين متوسطين)، وتحليل العامل (Factor Analysis)، ونحوها. وتضم النسخة «المهنية» للحزمة الإحصائية للعلوم الاجتماعية في نُسَخِها الأكثر حداثة (أي 20 وما فوق)، طرقاً عديدة لعملية التنقيب في البيانات، بما في ذلك نماذج الشبكات العصبية (Neural Network Models)، والطرق الآلية الخطية (Automated Network Models)، والتجميع (Clustering). وهذه الطرق برمتها سهلة الاستعمال، لأنها برامج (Programs) تعتمد «الإشارة والنقر»، ومدخلاتها (Inputs) ومخرجاتها (Outputs)، مصممة تصميماً محكماً. ولعل هذا، سيكون المكان الأفضل لمبتدئ ما، لتذوق بعض طرق التنقيب في البيانات.

وتضم حزمة التنقيب في البيانات الأكثر تقدماً، التي تدعى مُنْدمج الآي. بي. إم، والحزمة الإحصائية للعلوم الاجتماعية (IBM SPSS Modeler) اختياراً أكبر من طرق التنقيب في البيانات. ويعد هذا البرنامج أكثر تعقيداً للتعليم من الحزمة الإحصائية للعلوم الاجتماعية المطردة؛ لأنه يستلزم من المرء ترتيب أيقونات متعددة داخل عملية من العمليات، ووضع خيارات متنوعة، أو مَعْلمات (Parameters). ومع ذلك، يوفر المُنمذِج، مجموعة كاملة من أدوات التنقيب في البيانات.

وثمة منتوجات برمجية تجارية أخرى للحواسيب، تضم بعض أدوات التنقيب في البيانات داخل برمجياتها الإحصائية العامة، ومن ذلك، تقديم ماثووركس ماتلاب (MathWorks MATLAB) التنقيب في البيانات داخل «نظامي عدة» (Toolboxes) متخصصين: وهما الإحصاء والشبكات العصبية. وتضم حزمة ستاتيستيكا (Statistica) التابعة لـ«ستاتسوفت»، مجموعة من التنقيب في البيانات. وتعد تقنية الإكس. إل. ماينر (XL Miner) مضافاً تجارياً بالنسبة إلى التنقيب في البيانات التي تشتغل مع برنامج إكسيل (Excel) جدولي (Spreadsheet) لمايكروسوفت (Microsoft).

وبغض النظر عن البرمجية التجارية، ثمة حزمات مجانية متعددة من التنقيب في البيانات لفائدة الحواسيب؛ إذ تعد برمجيات الرابدماينر (RapidMiner)، مجموعة من البرامج الواسعة للتنقيب في البيانات، تم تطويرها في ألمانيا. ومؤخراً، ضمت معها برامج أخرى من برامج الويكا للتنقيب في البيانات (Weka DM)، مكتوبة في

اللغة «آر.»[3] (R Language). ونتيجة لذلك، يقدم الرابدماينر لحد الآن، أكبر عدداً من برنامج التنقيب في البيانات المتوفرة حالياً في منتوج برمجي مستقل. وهو أيضاً متوفر بالمجان على الرابط (http://rapid-i.com)، من أجل الاستزادة. وتأخذ البرمجيات وقتاً كبيراً قبل أن يتمكن الفرد من إتقانها؛ فهي تستخدم مقاربة مخطط انسيابي (Flowchart)، تشمل سحب الأيقونات إلى مساحة عمل، وربطها داخل برنامج أو تسلسل (Sequence). وهذه الفكرة مألوفة لدى واضعي برامج الحاسوب (Programmers)، ولكن قد تأخذ من الآخرين بعض الوقت لتعلمها. ومع ذلك، فالمستخدم لا يكتب أوامر أو شفرة (Code). إن ثمة قدراً كبيراً من التوثيق عبر الإنترنت، إلى جانب المدخل إلى برمجيات الرابدماينر الذي كتبه نورث (North 2012) وأفرد نسخة مجانية له على الرابط: http://dl.dropbox.com/ u31779972/DataMiningForTheMasses.pdf).

ويعد ويكا، أحد البرامج القديمة للتنقيب في البيانات، وهو متاح أيضاً بالمجان على الرابط (www.cs.waikato.ac.nz/ml/weka/). لقد تم تطويره في نيوزيلندا، وهو موثق توثيقاً جيداً بشكل استثنائي، بحيث يضم كتاباً موسوعياً (Witten, Eibe, and Hall 2011)، وبرامج تعليمية عبر الإنترنت: www.cs.ccsu. edu/~markov/weka-tutorial.pdf).

وإن راتل (Rattle) (http://rattle.togaware.com)، واجهة من واجهات المستخدم الرسومية (Graphical User Interface) المجانية بالنسبة إلى مجموعة من أدوات التنقيب في البيانات المتوافرة في لغة «آر.»، (و«آر.» نفسه تحميل مجاني). كما أن «راتل» موثق توثيقاً جيداً، بما في ذلك احتوائه على كتاب مدرسي (G. Williams 2011). ويعد «تراماينر» (TraMiner) (http://mephisto.unige.ch/ traminer) برنامجاً مجانياً من البرامج المتخصصة، التي تم تطويرها في سويسرا لتحليل متواليات وبيانات طولانية (Longitudinal). وليس هذا بديلاً، ولكن مكملاً بشكل أعم لبرمجيات التنقيب في البيانات.

---

(3) اقترن حرف «الراء» بكلمة «لغة» نسبة لحرف الراء الموجود في بداية الاسمين الأولين: روس إيهاكا (Ross Ihaka) وروبرت جانتلمان (Robert Gentleman)، لأن الفضل يرجع إليهما في اكتشاف هذه البرمجية (المترجم).

ولا أحد يعلم حزمة البرمجيات المتنافسة التي ستسود في الأعوام القادمة، ومن ثم، سيكون من الصعب علينا أن نوصي ببرمجية تستثمر فيها جهدك لتتعلمها. وإذا كانت تهمك سهولة الاستخدام أكثر من أي شيء آخر، فيمكنك - إذن - البدء بالحزمة الإحصائية للعلوم الاجتماعية المهنية (SPSS Professional)، أو «غامب» (JMP). ومن ناحية أخرى، إذا أردت الولوج إلى اللوحة الكاملة لتقنيات التنقيب في البيانات، فإن المُنمذِج (Modeler) أو «رابدماينر»، قد يكون اختياراً جيداً.

ملاحظة تحذيرية حول معدات الحاسوب

إن معدي برمجيات التنقيب في البيانات لأجهزة الحاسوب، يميلون إلى التقليل من أهمية تهيئة (Configuration) البرمجيات الضرورية لاستخدام منتوجاتهم بفاعلية. وقد دفعت برمجيات التنقيب في البيانات، أجهزت الحواسيب القائمة على «الويندوز» إلى حدودها القصوى؛ فعند استخدام أجهزة الحاسوب المكتبية العادية لتشغيل برمجيات التنقيب في البيانات، يكتشف المرء أن بعض التحليلات تشتغل ببطء شديد، وبعضها يصاب «بعطل»، أو «يتوقف فجأة»، حتى عندما تكون مجموعات البيانات غير كبيرة. ولتجنب تلك الإحباطات، من الأفضل استخدام جهاز حاسوب قوي ما أمكن، يحتوي - على الأقل - على 8 «جيغا بايت» (GB) من الرامات (RAM)، أي الذاكرة العشوائية في الهواتف والحواسيب، (ويفضل أن يكون أكثر من ذلك)، ومعالج معلومات متعدد النواة (Multicore Processor) (مثلاً معالجات «إنتل» من الجيل السادس، (Inteli 7)). وحتى ذلك الحين، قد تحتاج إلى استراحة لتناول قهوة، تاركاً في الوقت نفسه بعض البرامج تشتغل.

وتحتاج الكمية الكبيرة من المعلومات إلى محركات أقراص صلبة (Hard Drives)، ولكن أصبحت محركات تيرابايت -1 أو-2، خيارات غير مكلفة عند شراء حاسوب جديد. أما بالنسبة إلى معظم مجموعات البيانات (Datasets)، فتكفيها محركات أقراص صلبة صغرى. وتشكل - على ما يبدو - قراءة البيانات عقبة عندما يكون التنقيب في البيانات على جهاز الحاسوب، ولعل سرعة معالجة الذاكرة ووحدة المعالجة المركزية (CPU)، هي العوامل المحددة.

مصطلحات أساسية

يعتبر التنقيب في البيانات حقل معرفي متعدد التخصص، ساهم فيه كُلّ من علماء الحاسوب، والرياضيين، وعلماء الاجتماع التطبيقي. وتعكس مصطلحات التنقيب في البيانات هذه الأصول المتنوعة. هناك بعض المصطلحات الأساسية والمفاهيم التي ينبغي على القارئ الاطّلاع عليها منذ البداية.

- إن ما يصطلح عليه الإحصائيون بالمتغيرات (Variables) - مثلاً، طول شخص ما، ووزنه، ولون عينيه، أو عنوان عميل ما، ورقم هاتفه، ورمزه البريدي - هي عادة ما تُدعى سمات (Features) أو ميزات (Attributes) من لدن علماء التنقيب في البيانات، وعلماء الحاسوب.

- يميز علماء الإحصاء بين المتغيرات المستقلة (Independent Variables) (التي هي متنبئات (Predictors))، والمتغيرات التابعة (Dependent Variables) (وهي القياس الذي تم تنبؤه)، وعندما يتحدث علماء التنقيب في البيانات عن الشيء نفسه، سيشيرون إلى السمات أو الميزات التي تُتنبأ بهدف ما. وفي سياقات معينة، يستعملون أيضاً مصطلح فئة (Class) أو رقعة تعريف (Label) (عوض هدف)، قاصدين بذلك المتغير التابع المُتنبَّأ به.

- يحتوي نموذج (Model) ما، سمات أو ميزات رقمية، ممزوجة بطريقة رياضية داخل تنبؤ من تنبؤات متغير الهدف (Target Variable). وفي حالات عديدة، يعد نموذج من نماذج التنقيب في البيانات، معادلة تربط قيم سمات مرصودة عديدة بقيمة متنبأ بها بالنسبة إلى المتغير الهدف. وغالباً ما يتم بلوغ ذلك التنبؤ من خلال عملية ضرب القيمة المرصودة (Observed Values) لكل متغير أو سمة في عدد ما (**الوزن** أو **المعامل** (Coefficient)) خاص بذلك المتغير، وبعدها إضافة تلك المكوّنات معاً. وإن هذه القيم المناسبة لتلك الأوزان والمعاملات هي ما يُبتُّ فيها البرنامج (أو يستكشفه أو يتعلمه) لدى بناء نموذج ما.

- إن علماء التنقيب في البيانات، يتحدثون عن تركيب نموذج ما. وتشير هذه

العبارة أحياناً، إلى انتقاء تقنية نمذجة معينة؛ وأحياناً تشير إلى اختيار المتغيرات وشكلها ضمن نموذج، وتعديلاتها. وأحياناً أخرى، تشير العبارة إلى عملية ذات قيمة تقريبية مفرطة، حيث من خلالها يقترب نموذج ما - تدريجياً - من وصف البيانات وصفاً دقيقاً.

- تدعى بعض المقاييس (المدرجة في قسم لاحق)، علم الإحصاء التطابقي (Fit Statistics) أو حساب الدوال. إنها تصف مدى تطابق البيانات مع نموذج التنقيب في البيانات، أي إلى أي حدّ تُطابقُ القيمة المتوقعة لهدف ما بالنسبة إلى كُلّ حالة أو شخص، القيمة الحقيقية المرصودة لذلك الهدف بالنسبة إلى ذلك الشخص. إن هدف تحليل من تحليلات التنقيب في البيانات، إنتاج نموذج دقيق التنبؤ، أو كما نقول، يطابق البيانات بشكل جيد. ويمكن مقارنة الإحصاء التطابقي للبت في النموذج أو الطريقة التي تقوم بأداء جيد لمعالجة مجموعة بيانات محددة.

- ويشير مصطلح التعلم الآلي (Machine Learning) إلى تحليلات الحاسوب التي تنتج نموذجاً يتنبأ بأنماط في بيانات، أو يصنفها، أو يحددها. وإن العديد من طرق التنقيب في البيانات هي طرق تكرارية (Iterative)، إذ تمر في البداية، عبر سلسلة من الخطوات، التي تقدم تقديراً أولياً أو جواباً. وبعدها، تحصل بعض الطرق على تقديرات أفضل، من خلال إضافة مزيد من الأدلة (مزيد من الحالات أو البيانات) لتغيير التقديرات الأولى. وتعمل طرق أخرى بمبدأ التجربة والخطأ (Trial and Error)، إذ تُحدثُ تغييرات صغيرة على التقديرات الأولى، وترى ما إن كان التنبؤ المحصل عليه أفضل من التنبؤ السابق. وفي كلتا المقاربتين، يعيد برنامج التنقيب في البيانات، سلسلة من الخطوات مرات متعددة - أي تتكرر - حتى تصبح التقديرات أو الحلول أكثر دقة مع كُلّ دورة إضافية على حدة. وإن هذه العملية التدريجية، التي تشمل تقديرات أفضل على التوالي، تفضي إلى استعارة التعلم الآلي. يميز علماء التنقيب في البيانات بين التعليم الآلي الخاضع للإشراف والتعلم الآلي غير الخاضع للإشراف، وذلك لكون نوع التعليم الأول يشير إلى طرق

تلك البيانات حيث وجود كُلّ من المتغيرات المستقلة، والمتغيرات غير المستقلة على السواء (أي سمات وهدف ما أو رقعة تعريف (Label)). وفي مرحلة بناء النموذج، يدرك المحلل سلفاً، القيمة الحقيقية للهدف أو للمتغير المستقل بالنسبة إلى كُلّ حالة على حدة. ومن ثم، يضم النموذج استكشاف صيغة أو تعلمها، تتنبأ بشكل دقيق القيمة المرصودة للهدف، مستخدمة القيم المرصودة للسمات، ويدعى هذا أيضاً النموذج التعليمي (Training). ومن ناحية، «تشرف» البيانات المستهدفة على عملية التعلم (Learning Process). وفي مراحل متعاقبة من البحث، قد تستخدم تلك الصيغة أو ذلك النموذج للتنبؤ بقيم الهدف بالنسبة إلى بيانات جديدة، حيث القيم الحقيقية غير معروفة (وتدعى أحياناً بيانات خارج العينة (Out-of-Sample)). وفي المقابل، هناك طرق أخرى أو أدوات للتنقيب في البيانات حيث انعدام أي متغير هدف (أو رقعة تعريف أو فئة) يتنبأ به. وفي لغة علم الإحصاء، ليس هناك «متغير مستقل»، ويدعى هذا النوع الثاني من التنقيب في البيانات الذي يفتقر إلى الهدف، التعليم **غير الخاضع** للإشراف. ولا يزال برنامج الحاسوب أو نموذج الحاسوب في طور التعلم (إيجاد بنية)، ولكنه لا يستخدم متغير الهدف باعتباره مرشداً له. وما السعي إلى وجود تجميعات ذات حالات متشابهة داخل مجموعات بيانات إلا مثال واحد للتعليم غير الخاضع للإشراف.

- في مجال التنقيب في البيانات، يشير مصطلح اختيار السمات (Feature Selection) إلى تقليص عدد المتغيرات أو السمات ليتم تضمينها في نموذج من خلال تحديد المهم منها وسحب الباقي، بحيث يمكن - مع ذلك - لما تبقى منها التنبؤ بالهدف.

- ولاستخلاص السمات (Feature Extraction) الغاية نفسها، المتمثلة في بلوغ متغيرات أقل، غير أنه في استخلاص السمات، تُنتِج المتغيرات الأصلية المنهارة رياضياً، مجموعة محدودة جديدة من المتغيرات داخل متغيرات قليلة جديدة، من خلال مزج بعضٍ منها ضمن مقاييس (Scales).

- وأحياناً يدعى النمط أو البنية في البيانات، الإشارة. وبسبب خطأ مقياس (Measurement Error) أو تقلبات عشوائية (Random Fluctuations)، فإن هذه الإشارة تمتزج مع الضجيج (أو تتلوث به). ويأتي الضجيج من انعدام الدقة في القياس، أو من عوامل سياقية فريدة، تؤثر في حالات معينة أو أشخاص معينين في مجموعة البيانات (Dataset) على نحو مختلف عن حالات مماثلة أخرى. وعادة ما يتم تصور الضجيج باعتباره عشوائياً، بما أنه - من حيث التصور - نقيض الأنماط أو البنيات في البيانات. ويأتي هذا التماثل انطلاقاً من الأيام الأولى التي ظهر فيها جهاز الراديو، عندما كاد صوت (أي الإشارة) المذيع أن يلحقه تشويش بسبب الطقطقات وضجيج ما، ناتج عن خلفية أخرى، تجعل من الصعب إنتاج الإشارة. وستضم البيانات الأولية دائماً مزيجاً من الإشارة والضجيج، وتسعى كُلّ تحليلات التنقيب في البيانات إلى التمييز بين الإشارة والضجيج.

- وقد عمم مؤرخ العلوم - توماس كوهن (Thomas S. Kuhn) (1962) - مصطلح **النموذج** (النموذج الأصلي) (Paradigm) للإشارة إلى مدارس الفكر العلمي. وصور كوهن تقدم العلوم، باعتباره عملية تنافسية اصطدمت فيها أحياناً مدرسة من مدارس الفكر (نموذج واحد) - ذات باحثين، وطرق بحث خاصة بها - مع مدرسة أو نموذج جديد، ضم منخرطين، وتصورات، وطرق بحث مختلفة. وعندما يتفوق نموذج جديد على آخر قديم، يسمي كوهن ذلك النقلة النوعية (Paradigm Shift). وفي هذا الكتاب، سنقارن ما نسميه النموذج التقليدي أو الثابت لتحليل البيانات الكمية بالتنقيب في البيانات، التي تعتبره النموذج الأصلي ناشئ جديد. قد يحدث التنقيب في البيانات نقلة نوعية، ولكن من الممكن أيضاً أن تُستوعَب تقنيات التنقيب في البيانات ببساطة داخل نموذج تحليل البيانات القديمة في المستقبل. ويشير علماء التنقيب في البيانات إلى بُعديّة البيانات (Dimensionality)، للحديث مثلاً، عن مشكل ذي بعد مرتفع (High Dimension)، أو عن مشكل يشير إلى تخفيض الأبعاد (Dimension Reduction)، وقياس المساحة. كُلّ هذه المصطلحات تستعمل استعارة حيزية للتفكير في البيانات؛ فلنشرح، إذن، هذه الاستعارة.

توجد في الحيز المادي الذي نعيش فيه، ثلاثة أبعاد - الطول، الارتفاع، والعمق - ذات إحداثيات (Coordinates) ممثلة على المحاور x وy وz. ويمكن لكُلّ من هذه الأبعاد الموجودة في الحيز أو الفضاء، تمثيل متغير واحد في مجموعة بيانات ما. ومن ثم، فإذا كانت لدينا بيانات بشأن ثلاثة متغيرات - تخص مثلاً طول الأشخاص، ووزنهم، ومعدل دخلهم - فستتعامل مع متغير الطول بصفته x، ووزنه بصفته y، ومعدل دخله بصفته z. وبعدها، يكون بإمكاننا تخطيط(Plot) كل ملاحظة في هذا الحيز ذي الثلاثة أبعاد، وتحديد موقع القيم على المحاور x وy وz، لتمثيل طول كُلّ شخص على حدة، ووزنه، ومعدل دخله، ووضع نقطة (Dot) في الحيز الذي يوافق قيم x وy وz لذلك الشخص.

وإذا واصلنا تنقيط مجموعة البيانات برمتها، فسنرى آلاف النقاط في الحيز، بعضها موجود ضمن تجميعات كثيفة، وبعضها الآخر قائم بذاته. وإن هذه النقاط التي وضعت للأشخاص الذين يملكون قيماً مماثلة محددة على هذه المتغيرات أو الأبعاد الثلاثة، يتدانون فيما بينهم، في حين إن الأشخاص الذين يختلفون فيما بينهم وفق الأبعاد الثلاثة، يتباعدون على نحو مستقل.

ويمكن للرياضيين أن يصوروا أكثر من حيز بمئات أبعاده، ويصطلحون على تسميته بالحيز ذي الأبعاد المرتفعة (High-Dimensional Spaces)؛ ففي عالمنا ذي الأبعاد الثلاثة، لا يمكننا رسم حيز ذي الأبعاد المرتفعة أو بنائه، ولكن يمكننا تصور عالم له أبعاد عديدة. وهذا أمر مفيد، لأن مجموعات البيانات تضم - إجمالاً - أكثر بكثير من ثلاث متغيرات، وتوافق مجموعة بيانات ما، ذات متغيرات عديدة، حيزاً ذا أبعاد مرتفعة.

إن كُلّ ملاحظة في مجموعة بيانات ما، يمكن (في خيالنا) تنقيطها في نظام إحداثي (Coordinate System) ذي مئات الأبعاد، وليس فقط ثلاثة، بحيث يمثل كُلّ بعد متغيراً واحد. ويستخدم علماء التنقيب في البيانات حيز الاستعارة للحديث عن قياس المساحة، ويقصدون بذلك الحيز ذا الأبعاد المتعددة الذي يضم بياناتهم. كما يفكرون أيضاً في البنية داخل بياناتهم أو في العلاقات بين المتغيرات في البيانات من حيث الأنماط والأشكال في هذا الحيز النظري ذي الأبعاد المرتفعة.

وفي ضوء هذه الاستعارة، تضم بعض البنيات سحاباً كثيفاً من نقاط البيانات مجتمعة في هذا الحيز المتعدد الأبعاد، لأن قيمها في عدة متغيرات أو أبعاد متماثلة. ويتم تخيل بنيات أخرى باعتبارها نقاط بيانات منتظمة في خط طويل. ومع ذلك، إن تمثل بنيات أخرى (أو علاقات بين متغيرات) باعتبارها مستويات مسطحة، أو أسطح منحنية أو أسطح مشكلة تشكيلاً غريباً. (يسمي الرياضيون هذه الأشكال تحدبات (Manifolds))، بحيث يمثل كُلّ شكل من الأشكال، علاقة رياضية ما، بين بعض المتغيرات في مجموعة البيانات.

إن بعض طرق التنقيب في البيانات - في هذا العالم التصوري لأبعاد عديدة - تشتغل وفق عملية إسقاط، تُترجم البيانات رياضياً من حيز ذي أبعاد أكثر ارتفاعاً إلى حيز ذي أبعاد أكثر انخفاضاً، لأنه من السهل التعامل مع مسألة رياضية ذات أبعاد أقل. إن هذا الإسقاط ممكن، لأن البنيات أو العلاقات البارزة في الحيز ذي الأبعاد الأكثر ارتفاعاً، غالباً ما تكون محفوظة عندما يتم إسقاطها في حيز ذي بعد أقل انخفاضاً. وهذا يعادل عملية تقليص متغيرات عديدة، واكتشاف أن العلاقات الأساسية محفوظة.

وأما طرق أخرى من طرق التنقيب في البيانات، فتعمل في الاتّجاه المعاكس: إذ إن المشكلة التي يصعب حلها بسهولة في حيز ذي أبعاد أكثر انخفاض لدى إسقاطه على حيز ذي أبعاد أكثر ارتفاعاً، قد تصبح معالجته أسهل رياضياً باستخدام حيلة النواة (Kernel Trick). وتستخدم طرق عديدة من طرق التنقيب في البيانات هذه الاستراتيجية، من أجل تصنيف الملاحظات، كما ستبين الأمثلة ذلك لاحقاً.

الفصل الثاني

عقد المقارنات بين نموذج التنقيب في البيانات وبين المنهجية الإحصائية التقليدية

يقدّم التنقيب في البيانات منهجية لتحليل البيانات، تختلف في مناحٍ مهمة عن الطرق الإحصائية التقليدية التي بَسَطَتْ هيمنتها خلال العقود القليلة الماضية. في هذا القسم، سنبرز بعض التباينات بين النموذج الأصلي (براديغم)[1] الناشئ للتنقيب في البيانات، وبين المقاربة الإحصائية التقليدية لتحليل البيانات قبل تفصيل القول - ضمن فصول لاحقة - في الطرق أو الأدوات الفردية التي تشكل التنقيب في البيانات. ولبيان هذه التباينات، سنتوسل بالانحدار المتعدد[2] (Multiple Regression)، للإشارة إلى المنهج التقليدي، بما أن هذه الطريقة الإحصائية تشكل دعامة تحليل البيانات التقليدية في العقود الأخيرة - إلى جانب امتداداتها وفروعها، بما في ذلك الانحدار اللوجيستي، وتحليل الحدث التاريخي، والنماذج متعددة المستويات، ونماذج التسجيل الخطي (Log-Linear Models)، ونمذجة المعادلة الهيكلية (Structural Equation Modeling).

وستبرز هذه المقارنة المنهجية بعض مواطن الضعف والصعوبات داخل النموذج الأصلي التقليدي، التي لم تعد إشكالية في منهجية التنقيب في البيانات.

(1) لم استعمل كلمة «نموذج» بمفردها بل أضفت لفظ براديغم إلى جانبها، درءاً لأي خلط قد يحصل في الفهم بين «Model» و«Paradigm»، الكلمتين الإنجليزيتين اللتين تترجمان بنفس اللفظة العربية «نموذج» (المترجم).

(2) تترجم العبارة أيضاً بـ «التراجع المضاعف»، وقد تكون الأنسب في المجال الحاسوبي، غير أننا لا نمانع استخدام الاثنين (المترجم).

ومع ذلك، لا يعني خلو هذه المنهجية من المشاكل، عندما ترانا نشدد على مزالق النموذج التقليدي، بل على العكس من ذلك تماماً؛ فللتنقيب في البيانات نقائصه، التي سيحدد بعض منها في الأقسام اللاحقة.

القوة التنبؤية في النموذج الإحصائي التقليدي

في التحليلات الإحصائية التقليدية مثل الانحدار، يركز محلل ما - عادة - على القيم الرقمية، أو معامِلات (Coefficients) ذات متنبئات مهمة في نموذج ما. إنَّ القوة التنبؤية أو التناسب التنبؤي (Fit) لذاك النموذج، عادة ما تكون له أهمية ثانوية (Breiman, 2001). وكل ذلك راجع إلى الهدف الرئيس للعديد من الباحثين ممن يستخدمون الطرق التقليدية، اختبار فرضيات حول متنبئات (Predictors) خاصة، أو فهم كيفية ارتباط المتنبئات الفردية بالمتغير التابع (Dependent Variable). وتمثل تلك العلاقات، المعامِلات بالنسبة إلى كُلّ متغير في انحدار من الانحدارات أو نموذج تنبؤي آخر.

ومع ذلك، دائماً ما تُذكر قياسات تَناسُب النموذج (Model Fit)، في تحليلات بيانات تقليدية. ويعد قياس R^2، وقياس R^2 المعدل (Adjusted)، أكثر القياسات شيوعاً، إذ عادة ما يتم تفسيرها بنسبة تباين المتغير التابع، الذي يُشرح بمزج التنبؤات في النموذج. وتوجد قياسات أكثر تعقيداً للتناسب في سياقات أخرى، ومجموعة كاملة من إحصاءات التناسب، بما فيها A' (A عدد أولي)، و«كابا» (Kappa)، ومعيار أكايكي للمعلومة (AIC)، ومعيار بايز للمعلومة (BIC)، ومعيار المخاطرة في التضخم (RIC)، ومعيار بايز الممتد للمعلومة، ومقاييس شبه-R^2، واحتمال تسجيل نسبة 2- (Log-Likelihood 2-). ولكن الفكرة العامة التي نحن على وشك توضيحها، تنطبق على هذه المقاييس كلها.

وفي مقالات تستخدم الطرق التقليدية، تُنشر في مجلات بحث رائدة في العلوم الإنسانية، عادة ما توجد نماذج تنبؤية حيث نسبة التباين التي تم تفسيرها فيها، جداً متواضعة، قد تصل مثلاً إلى ٪25 أو أقل من ذلك. ولكن هذا المستوى المنخفض من القوة التفسيرية، نادراً ما ينظر إليه على أنه ينال من مصداقية دراسة ما، أو يتم التعامل معه باعتباره اتّهاماً لجودة النموذج.

ومن النادر أيضاً ما يركز كتاب مقالات بحثية في مجلات عديدة، على مقدار التباين في متغيرهم التابع الذي يتم تفسيره من قبل نموذجهم الخاص؛ بل من النادر جداً، وجود أي تفسير موضوعي حول التباين غير المفسر لنموذج ما. حدث استثناء واحد منذ عقود مضت عندما حاد كريستوفر جينكس (Christopher Jencks) وزملاؤه عن العادة في كتابهم الرائد **اللا مساواة** (*Inequality*) (1972)، وفسروا التباين غير المفسر لنموذجهم الخاص للحركية الاجتماعية بتأثير «الحظ». وخلف ذلك الكثير من الانتقادات (Coleman et al., 1973).

ونتج عن هذا الجدال – على ما يبدو – إجماعاً داخل النموذج الأصلي التقليدي يفيد بوجود اعتبار التباين غير المفسر (Unexplained Variance)، نابعاً من مزيج خطأ مقياس (Measurement Error) وعوامل سببية محذوفة. وما دام انحدار ما أو نموذج آخر ذو دلالة إحصائية بشكل عام، وتوجد متنبئات فردية ذات دلالة إحصائية داخل النموذج، فإن الإعلان عن نموذج تظل فيه الغالبية العظمى من التباين غير مفسرة، يبقى مقبولاً باعتباره طبيعياً ومناسباً لدى العديد من الباحثين، والمجلات الرئيسة في العلوم الاجتماعية والسلوكية.

وفي المقابل، يركز التنقيب في البيانات – على نحو أقوى بكثير – على تعظيم القوة التنبؤية لنموذج ما، مما يعني تقليص مقدار التباين غير المفسر قدر الإمكان. وإن تفسير 25٪ من تباين المتغير التابع فقط، قد يعتبر أمراً غير ملائم من قبل العديد من علماء التنقيب في البيانات. وكما سنأتي على ذلك لاحقاً، سيستكشف عالم من علماء التنقيب في البيانات طرقاً مختلفة – وأحياناً يمزج العديد منها – وذلك تحديداً لتعظيم القوة التنبؤية العامة. ويقوم علماء التنقيب في البيانات بذلك، لأن التنبؤ الدقيق هو في الغالب غايتهم الرئيسة في النمذجة، بما أن القيم المتنبأ بها، ستستخدم في حالات العالم الواقعي للإفصاح عن قرارات وإجراءات.

وخلاصة القول، إن المنهجية الإحصائية التقليدية، تركز على المعامِلات الفردية بالنسبة إلى المتنبئات، ولا تكترث كثيراً للقوة التنبؤية. ويعمل التنقيب في البيانات العكس، وهذا التباين في الأهداف، يشكل النقطة الرئيسة الأولى للتباين الحاصل بين التنقيب في البيانات، والنموذج الأصلي التقليدي.

يا تُرى، ماذا يمكن أن يقول التنقيب في البيانات والإحصاءات التقليدية لبعضهما بعضاً إذا كانت لديهما أهداف مختلفة؟ ولما كان تركيز التنقيب في البيانات يقع على القوة التنبؤية (Predictive Power)، تمكنت بذلك من تطوير بعض الأدوات التحليلية الجديدة القوية؛ ولكن ليس من الواضح دوماً، مدى إمكانية اندماج نقاط قوة التنقيب في البيانات في التنبؤ، ضمن إطار العلوم الإنسانية التقليدية التي تولي أولوية خاصة لتقييم فرضيات حول متنبئات خاصة، وتأثيراتها التقديرية. من المرجح - في رأينا - أن يحدث التنقيب في البيانات تغييرات كبرى في مجال البحث الاجتماعي والسلوكي، وفي الغاية الإحصائية في البحث الطبي الحيوي (Biomedical). وفي كثير من الحالات، تقدم أدوات التنقيب في البيانات، قدراً من القوة التفسيرية تفوق بكثير النماذج الإحصائية التقليدية التي من الأرجح، ينجذب الباحثين إلى استخدامها. ولكن تركيز العلماء الاجتماع، والسلوكيين، وباحثين آخرين على فهم آليات سببية (Causal Mechanisms)، والأهمية التي يولونها لتقديرات التأثيرات بالنسبة إلى المتنبئات الفردية (تقاس باعتبارها معامِلات متغيرات محددة)، لا تختفي على الأرجح. وتظهر إحدى التسويات في تطوير بعض أدوات التنقيب في البيانات الجديدة التي توفر معلومات حول الآليات، إضافة إلى الانشغال القديم للتنقيب في البيانات، بتعظيم الدقة في التنبؤ (انظر مثلاً، Pearl, 2000).

اختبار الفرضية في المنهجية التقليدية

لقد تمت داخل النموذج الأصلي الإحصائي التقليدي الذي هيمن على الطرق الكمية، والصلات (Linkages) بين النظرية وتحليل البيانات من خلال اختبار فرضيات حول معامِلات إحدى المتغيرات التابعة أو أكثر، في نموذج تنبؤي ما. على سبيل المثال، قد يركز باحث ما أو محلل بيانات على مسألة ما، إن كان معامِل انحدار ما، بالنسبة إلى متنبئ محدد ومهم نظرياً، له دلالة إحصائية؛ ففي مُخرَج الانحدار، تتم عملية نقل معامل كُلّ متنبئ إلى جانب إحصائية اختبار (اختبار - ت(t-test))أو اختبار - ز (z-Test)، وقيمتها p[3] (p-value)، المترابطة، أو مستوى الدلالة

(3) يشير الحرف اللاتيني p إلى الكلمة اللاتينية «probare»، ويعني القيمة الاحتمالية. وهو مصطلح يستعمل في مجال الإحصاء، أي أنه عبارة عن عدد يستعمل في تأويل أو تقييم المقاييس الإحصائية (المترجم).

(Significance Level). والقيمة p المترابطة بكل متنبِّئٍ (Predictor)، هي احتمال الحصول على قيمة احصائية الاختبار التي تعد كبيرة مثل تلك التي رصدت، ومتوقفة على صدق الفرضية الصفرية أو (العدم)[4] (Null Hypothesis).

ونادراً ما يتم اختبار فرضية ما - داخل المنهجية التقليدية - التي تفحص إمكانية أن يكون للتأثير المشترك لمتغيرات عديدة دلالة إحصائية. وأحياناً، يتم اختبار فرضية ما، لاستكشاف ما إن كان نموذج واحد عموماً، مختلفاً - بشكل كبير - عن نموذج بديل أو متفوقاً عليه.

وبعيداً عن هذه التفاصيل، يقدم اختبار الدلالة (Significance Testing) داخل المنهجية التقليدية، طريقة من طرق الحُكم على إمكانية أن تكون نتيجة ما، تمثيلية (Representative): أي ما إن كانت القيمة التقديرية (An Estimate)، المشتقة من عينة واحدة أو من مجموعة من الترصدات، ستكشف عن دقتها لدى تطبيقها على عدد أكبر من السكان الذين أخذت منهم العينة (Sample). وعندما نجد لمعامل الانحدار «دلالة، إحصائياً، نستنتج عدم إمكانية حدوث قيمة معينة - رصدناها في عينتنا - بمحض الصدفة عبر الخطأ العيني»[5] (Sampling Error). يعد اختبار الدلالة، إذن، طريقة من طرق تقييم إمكانية أن تنطبق نتيجة ما في عينة شخص ما، على العدد الأكبر من السكّان التي أخذت منها العينة.

ومع ذلك، أثار العديد من علماء الإحصاء انتقادات خطيرة بشأن الممارسات المتفق عليها بشكل عام في العلوم الاجتماعية والسلوكية، وفي البحث الطبي بما فيه اختبار الدلالة الذي أصبح يعرف بـ «اختبار الدلالة الجدلي» Significance Test) (Controversy) (Morrison and Henkel 1970). يرى هؤلاء النقاد أن العديد من الباحثين، يسيئون استخدام اختبارات الدلالة على نحو يقوض صلاحية النتائج الواردة في تقريرهم. وسنلخص بعض انتقاداتهم، ثم نبين أن التنقيب في البيانات دائماً ما يتبنى منهجية بديلة لتقييم النتائج، منهجية لا تعتمد اعتماداً كبيراً على اختبار

(4) تترجم عبارة «Null Hypothesis» في عالم المال بـ «فرض باطل»، وفي مجال الطب بـ «فرضية البطلان».

(5) تترجم العبارة أيضاً بـ «خطأ المعاينة» (المترجم).

الدلالة التقليدي. ومن ثم، يتجنب التنقيب في البيانات العديد من سوء الاستخدام الطويل الأمد، لاختبار الدلالة.

وتؤكد إحدى الانتقادات أن قرار تجاهل متنبأ ما ذي قيمة *p* (دلالة) 051.، واعتبار متنبأ ذي قيمة p 094. ذي دلالة، مضلل. وتفيد إحدى مضامين ذلك، بضرورة تركيز النتيجة أو حجم التأثير (Effect Sizes) - حجم المعامل أو التأثير - على الجانب التحليلي أو التأويلي، بدلاً من التركيز فقط على ما إذا كان معامل المتنبأ مهماً أو غير مهم إحصائياً (Nickerson 2000).

أما المأخذ الثاني على سوء استخدام اختبار الدلالة، فيتجلى في كون القيمة الحرجة المستخدمة بشكل مشترك من قبل باحثين، للبت في إمكانية دلالة معامل ما إحصائياً، صغيرة جداً في سياقات عديدة، ويؤدي إلى انتشار أخطاء من نوع (Type 1 I Errors (النتائج إيجابية كاذبة (False Positives)). ويشمل هذا الجدال الدائر، مخاطر التّعدد (Multiplicity) عندما يضم الانحدار أو نماذج أخرى، العديد من المتنبئات (Benjamini 2010, Hsu 1996; Saville 1990; Schaffer 1995; Tukey 1991). وفي نماذج تضم العديد من المتنبئات، يرى النقاد عدم ملاءمة استخدام القيمة الحرجة التقليدية (Conventional Critical value) (قيمة T أو Z= 1.96) لتقييم الدلالة الإحصائية لكُلّ متنبئ، بما أن تلك القيمة الحرجة التقليدية تنطبق بشكل مناسب على مقارنة واحدة، وليس على متنبئات متعددة، كُلّ بحسب اختباره الدلالي (Larzelere and Mulaik 1977). ويزداد احتمال وجود نتيجة ذات دلالة إحصائية بالدخول إلى المتنبئات في أي نموذج، لأن واحدة من هذه المتنبئات - ببساطة - تكرر اختبار الدلالة العديد من المرات. وسيكون لمتنبئ واحد من أصل عشرين، دلالة عند $p \leq 0.05$، عبر الحظ (احتمال) فقط. ويدور هذا الجدال أيضاً حول إمكانية أن تكون الاختبارات بالنسبة إلى المتنبئات المختلفة في نموذج ما، مستقلة عن بعضها بعضاً بشكل حقيقي.

ويمكن تجنب مشكل التعدد أو عملية اختبار العديد من التأثيرات أو المعامِلات في نموذج واحد، من خلال تعديل القيمة الحرجة المستعملة للبت في المتنبئات ذات الدلالة إحصائياً للأخذ بعين الاعتبار عدد المقارنات المتعددة. ويتجلى أحد

الحلول المتحفظة (Conservative) إحصائياً، في استخدام تصحيح بونفيروني (Bonferroni Correction) للمقارنات المتعددة. وإذا كانت هناك خمسة متنبئات - مثلاً - عوض القبول بأي متنبئ كان، بحيث تكون p<.05 قيمة احتمالية ذات دلالية، فإن المرء يقبل فقط متنبئاً من المتنبئات، بحيث تكون p<.01 (أي قسمة القيمة التقليدية لـ 0.05 على عدد المتنبئات). وهذا يعادل استخدام قيمة حرجة (2.58، عوض 1.96 بالنسبة إلى نموذج انحدار يضم خمسة متنبئات، أو قيمة 3.48 بالنسبة إلى نموذج ما، ذي مائة متنبئ).

وما تعديل بونفيروني (Bonferroni Adjustment)، إلا تصحيح تعدد ممكن، يمكن تطبيقه على اختبارات الدلالة التي تنقلها البرمجيات الإحصائية العادية. وتضم أكثر المنهجيات تطوراً للتعدد، حساب معدلات اكتشاف كاذبة (False Discovery Rates) ومعدلات الخطأ حسب العائلة (Family-Wise Error Rates) (Benjamini 2010). وإذا كانت المجلات البحثية تشترط هذه التعديلات بالنسبة إلى هذه النماذج التنبؤية التي تضم هذه المتنبئات المتعددة، فإن ورود النوع الأول من الخطأ (Type I Errors) (الإيجابي الكاذب) (False Positives)، سيتقلص بشكل كبير. ومع ذلك، تستمر المجلات البحثية البارزة في قبول استخدام قيمة حرجة لـ 1.96، في نماذج ذات متنبئات متعددة، مما يؤدي إلى تعريات حول عدم قابلية استنساخ البحوث (Ioannidis 2005)، على الرغم من عقود من الانتقادات في هذه الاتّجاهات.

وتتفاقم هذه القضية المترابطة باختبار الدلالة في سياق المتنبئات المتعددة عندما يبحث بعض الدارسين بشكل فعال عن تأثيرات ذات دلالة إحصائياً، من خلال تحليل متنبئات عديدة إلى أن يعثروا على نتيجة ذات قيمة T أو Z لـ 1.96 أو أكبر، ثم يضمونها في نموذج نهائي، ومقرر باعتبارها قيمة ذات دلالة. وتعد قيمة حرجة ما لـ 1.96 بالنسبة إلى اختبار الدلالة مضللة جداً، إذا ما تم تقييم مئات المتنبئات أولاً، قبل نقل فقط تلك التي أثبتت أن لها دلالة إحصائياً.

وزادت حدّة هذه المشاكل أكثر في البحث الطبي وفي تحليلات سلسلات جينية، بحيث أصبح خضوع آلاف اختبارات الدلالة للتجربة، أمراً شائعاً بشكل متزايد قبل نقل أهمها (Benjamini, 2010). وتحذر الكتب المدرسية التي تتناول الطرق

(Methods)، من البحث في المتنبئات الدلالية التي تعرف على نحو غير رسمي باسم «التنقيب» (Fishing) (التي تحمل معنى البحث والتنقيب)، أو «تجريف البيانات» (Data Dredging). كما توصي الكتب المدرسية، الباحثين بطرح فرضياتهم مقدماً قبل تحليل البيانات بغية تجنب إغراء المضي في البحث بعد المعلومة (Fact) عبر المتنبئات العديدة الممكنة. ولسوء الحظ، ما يزال العديد من الباحثين «ينقبون» عن النتائج ذات الدلالة، مستخدمين قيمة حرجة منخفضة (T:1.96 أو $p < 0.5$) بالنسبة إلى الدلالة الإحصائية.

وكما أشرنا إلى ذلك سلفاً، ثمة حلول ناجعة للتعامل مع التعدد داخل النموذج الأصلي التقليدي؛ ولكن معظم طرق التنقيب في البيانات - كما سنفصل القول في ذلك لاحقاً - تتبنى منهجية مختلفة لتقييم تمثيلية نموذج ما (مستخدمين شكلاً من أشكال المضاعفة (Replication))، المعروفة بالصلاحية المتبادلة (Cross-Validation)، تتحاشى مشكلة التعدد برمته، ولا يقوم على اختبار الدلالة. وتلكم هي النقطة الثانية الرئيسة التي يختلف فيها التنقيب في البيانات مع النموذج الأصلي التقليدي.

عدم تجانس التباين باعتباره تهديداً للصلاحية في النمذجة التقليدية

بعيداً عن مسألة التعدد، تتأثر أيضاً دقة الافتراضات الدلالية في النموذج الأصلي التقليدي ببعض الافتراضات النظرية الإحصائية التي تشكل الأساس لنماذج انحدار متعددة، وأقربائها الإحصائية (Berry 1993). سنوضح بعضاً منها بهدف وضع الأسس لمفارقة أخرى بينها وبين التنقيب في البيانات.

يقوم نموذج ما - لكُلّ حالة أو رصد على حدة، في مجموعة بيانات - بتقدير قيمة متنَبَّأة للهدف (متغير تابع). وإذا ما طرحنا هذه القيمة المتنَبَّأة من القيمة المرصودة، فسنحصل على عدد يعرف بالقيمة المتبقية (Residual)، التي تمثل خطأ التنبؤ (Prediction Error) بالنسبة إلى كُلّ رصد فردي. فالقيمة المتبقية إذن، نوع خاص من متغير من المتغيرات. وتلخيصاً للقيم المتبقية (أو الأخطاء) عبر كُلّ الترصدات، نقول إنها تحتوي على التباين غير المفسر لنموذج تنبؤي.

إن مجموعة من الافتراضات الكامنة وراء المنطق الإحصائي للانحدار المتعدد والطرق المتصلة به، تفيد بضرورة توزيع القيم المتبقية بشكل عادي، ذات تباين ثابت، ومتوسط قيمة الصفر (A Mean of Zero)، واستقلالية عن بعضها بعضاً. وعندما تكون هذه الافتراضات دقيقة، يقال عن الأخطاء إنها هوموسيداستيك (Homoscedastic)، وهي مصطلح يوناني يعني «ذات تباينات متساوية» أو متماثلة التفاوت.

وعندما تكون هذه الافتراضات غير دقيقة، يقال إنها هيتروسيداستيك (Heteroscedastic)، أي ذات تباينات غير متساوية. ويحدث عدم تجانس التباين (Heteroscedasticity) في الغالب، عندما تكون القيم المتبقية أو أخطاء التنبؤ أكثر انتشاراً (ذات تباين أعلى) بقيم منخفضة أو عالية لمتنبأ معين (أو X) من أخرى بقيم معتدلة لذلك المتنبأ X؛ أو أحياناً تكون القيم المتبقية أكثر انتشاراً بقيم منخفضة أو عالية (للمتغير التابع) Y من غيرها بقيم معتدلة للمتغير التابع Y. ومؤدى ذلك، أنّ النموذج التنبؤي دقيق (وله قيم متبقية صغيرة) عبر مجموعة معينة من قيم X أو Y، ويأخذ في التراجع (أي يصبح أقل دقة) في مكان آخر، ليبلغ في الغالب، قيماً قصوى من قيم X و/ أو Y.

ثمة اختبارات إحصائية لتحديد إمكانية أن تكون الأخطاء، ذات تباينات غير متساوية، ولكن غالباً ما سيرسم الباحثون، القيم المتبقية مقابل كُلّ متغير X و/ أو كُلّ متغير Y. وضمن هذه الرسومات البيانية، يأخذ عدم تجانس التباين شكل القِمع، بحيث يزداد التباين في الأخطاء على مستوى الجزء الكبير من القِمع.

ولماذا نخوض بشكل عميق جداً، فيما قد يبدو تفصيلاً تقنياً؟ **أولاً**: يعد عدم تجانس التباين مشكلاً، يظهر في تحليلات كمية عديدة. **ثانياً**: يخلف عدم تجانس التباين عواقب وخيمة تتسبب في التحيزات التي - في نظرنا - تقوض دقة بعض البحوث. **ثالثاً**: يقدم التنقيب في البيانات عدداً من السبل لمعالجة مشكل عدم تجانس التباين، أو التحايل عليه في بعض الأحيان.

ثمة أسباب متعددة تؤدي إلى حدوث عدم تجانس التباين، فتجعل منه مشكلة واسعة الانتشار.

• عندما تكون وحدات التحليل في مجموعة بيانات، تجمعات أو تراكمات (Aggregates) ذات أحجام مختلفة (مثلاً، احتواء مدارس على أعداد مختلفة من الطلبة)، وتكون متغيرات كُلّ وحدة على حدة (مثلاً)، متوسط درجات اختبار التقييم المدرسي (SAT) لدى الطلاب، يظهر في الغالب، عدم تجانس التباين، لأنه ستكون أخطاء أكثر في القياس بالنسبة إلى المدارس الصغيرة، حيث متوسط درجات اختبار التقييم المدرسي يقوم على عدد أكبر من الطلاب.

• ويحدث عدم تجانس التباين أيضاً، عندما تضم مجموعة بيانات ساكنات فرعية، التي تظهر علاقات مختلفة بين X و Y. .يكون تحليل ما، وهو يتنبأ بتأثير أخذ دروس التقوية داخل الكلية في ترك الدراسة، ذا تباين غير متساوٍ، وينتج معامِلات مضللة إذا ما ضمت العينة - مثلاً - طلبة المرحلة الجامعية من المجتمع، ومن الكليات ذات الزمن الممتد لأربع سنوات على حدّ سواء، وحدث أن كان لأخذ دروس التقوية في الكلية تأثير مختلف جداً في ترك الدراسة بالنسبة إلى طلبة كلية المجتمع (Community College)، عن طلبة الكلية ذات التكوين الممتد لأربع سنوات.

• كما يمكن أيضاً حدوث عدم تجانس التباين عندما تقاس المتنبئات بشكل غير مناسب، مثلاً عندما يستخدم الدخل عوض سجل الدخل متنبِّئاً (Predictor).

• **وأخيراً**، يحدث عدم تجانس التباين عندما تكون علاقة ما «ضرورية، ولكن غير كافية». على سبيل المثال، قد ترتفع نفقات إجازة السفر بارتفاع دخل الأسرة، بما أن المرء يحتاج إلى دخل كافٍ لتحمل تكاليف السفر. ولكن لا يستلزم ذلك ارتباط الدخل المرتفع بكثرة نسبة الأسفار. وبناء على ذلك، عندما يتم تنبؤ نفقات السفر انطلاقاً من دخل العائلة، يحدث قدر كبير من أخطاء التنبؤ (Prediction Errors) (أي القيمة المتبقية) على مستوى الدخل المرتفع أكثر مما يحدث على مستوى الدخل المنخفض. وسيظهر ذلك باعتباره علاقة إيجابية ما بين الدخل والفترة المتبقية (Residual Term).

إن عدم تجانس التباين منتشر في كُلّ أنواع البيانات، وله عواقب وخيمة على النماذج التقليدية؛ ففي حالة انحدار المربعات الصغرى العادية (Ordinary Least Squares Regression)، لا ينحاز عدم تجانس التباين في تقديرات معاملات

الانحدار أو المتنبِّئات، ولكنه ينحاز في تقديراته للأخطاء المعيارية بالنسبة إلى تلك المتنبئات، ومن ثم، يقدر اختبارات الدلالة تقديراً منحازاً، يتم نقله إلى تلك المعاملات من معاملات الانحدار (Regression Coefficients). وهذا ما قد ينتج خطأ من نوع I، مما يؤدي بالباحثين إلى استخلاص خاطئ مفاده أن لمعامل متنبِّئ ما، دلالة إحصائية، في وقت تنعدم فيه هذه الدلالة أصلاً - أو تنتج تحيزات متزايدة لأخطاء معيارية، تفضي إلى خطأ من نوع II، أي تؤدي بالباحثين إلى الاعتقاد في أن بعض المعاملات ليست لها دلالة، في وقت تتحقق فيه هذه الدلالة. والمشكلتان كلاهما تشكلان تهديداً للنمذجة التقليدية.

وفي حالة الانحدار اللوجيستي، والاحتمالية (Probit)، والتقنيات ذات الصلة، التي تتنبأ بالمتغيرات الثنائية أو القطعية، يكون لعدم تجانس التباين نتائج أكثر سوء، ذلك بأنه يتحيز لمعامِلات الانحدار، وكذا أخطائها المعيارية (R. Williams, 2010).

ولا يتفق كُلّ الباحثين مع حجّتنا التي تفيد بأن عدم تجانس التباين يخلق مشكلاً خطيراً للنماذج التقليدية؛ فالمختصون في علم الاقتصاد القياسي (Econometricians) - مثلًا - طوروا مقدرين (Estimators) خاصين للأخطاء المعيارية، تعرف بمقدري الشطائر (Sandwich Estimators)، أو مقدري هابير- وايت (Huber-White Estimators)، والأخطاء المعيارية القوية (Robust Standard Errors)، التي يقال عنها إنها تقلص من الانحيازات الناجمة عن عدم تجانس التباين. ولكن يشكك علماء إحصاء آخرين، في مصداقية هذه التدابير التصحيحية (Freedman, 2006)، لكونها لا تشكل حلاً سريعاً.

إن التنقيب في البيانات يقدم أدوات متعددة لتحديد و/ أو معالجة عدم التجانس (Heterogeneity) الذي يعزى حدوثه جزئياً إلى قياس متنبأ ما قياساً هزيلاً، أو إلى علاقته غير الخطية بالمتغير التابع. وسنعرض في قسم لاحق، لأدوات التنقيب في البيانات المعروفة باسم توزيع الخانات (Binning) أو تفريد أنتروبي (Entropy Discretization)، التي تسمح للمحللين بتحديد التأثيرات اللا خطية؛ وإذا ما تم تقدير هذه الأدوات بشكل صحيح في نموذج ما، فإن مصدر عدم تجانس التباين قد يزول.

يعد عدم تجانس التباين أحياناً، نتيجة لتفاعلات إحصائية مهمة، استُبعدت من نموذج ما. وتقدّم أدوات التنقيب في البيانات بما في ذلك تقسيم البيانات (Data Partitioning) أو أشجار القرار (Decision Trees) لصقل مجموعة بيانات للتفاعلات، وتحديد التفاعلات الأكثر أهمية. وإذا ما حدد المحللون تلك التفاعلات ونمذجوها، فإن عدم تجانس التباين سيتقلص. وعلى نحو مشابه، يوفر التنقيب في البيانات طرقاً من أجل تحديد مجموعات فرعية في مجموعة بيانات ما، ذات علاقات مختلفة بين المتنبئات والمتغير التابع. وبمجرد تحديد مجموعات فرعية مميزة، داخل مجموعة بيانات من خلال استخدام التنقيب في البيانات، يكون بإمكان الباحثين البتّ في إضافة إجراء تحاليل منفصلة لكُلّ مجموعة فرعية على حدة (Melamed, Breiger, and Schoon 2013). وفي كلا الحالتين، لا بُدَّ من أن يقلص هذا عدم تجانس التباين.

ولكن في حالات أخرى، يرى محللون أن فترة خطأ (Error Term) نموذجهم هي فترة خطأ ذات تباينات غير متساوية. ومع ذلك، عجزوا عن تحديد أسباب مشكلهم، على الرغم من جهودهم المثلى. وفي هذه الحالة - كما سنشرح ذلك في قسم لاحق - يملك التنقيب في البيانات «حيلة» لتعيير نماذج لم تُعيَّر في البداية بسبب عدم تجانس التباين. ولا تحدد هذه الحيلة أسباب المشكلة في المقام الأول، ولكن يمكن أن تقدم حلاً ناجعاً لإزالة تلك المشكلة.

وأخيراً، تعد العديد من طرق التنقيب في البيانات، لا معلمية (Nonparametric): ذلك بأنها لا تستلزم أنواع الافتراضات الإحصائية حول توزيع فترات الخطأ التي تقوم على مجموعة من الطرق التقليدية المنمذجة. وبينما تعجز طرق التنقيب في البيانات - في هذه الحالات - عن منع حدوث عدم تجانس التباين في البيانات، تستطيع مع كُلّ هذا، التحايل على بعض آثارها المدمرة أو الصعبة.

تحدي العينات المعقدة وغير العشوائية

في النموذج الأصلي التقليدي، عادة ما تقاس الاختبارات الدلالية الإحصائية لكُلّ متنبِّئ في نموذج انحدار ما، برزم إحصائية من خلال استخدام صيغ تفترض فكرة تشكيل البيانات، عينة عشوائية بسيطة (Simple Random Sample)، مستمدة من سكّان أكبر،

وأحياناً، يكون ذلك الافتراض غير مبرر. وتشمل العديد من الدراسات الاستقصائية، مخططات أخذ العينات[6] متعددة المراحل.

أولاً: أخذ العينات العشوائية بين وحدات ذات مستوى أعلى مثل المدن أو الرموز البريدية (Zip Codes)، وبعدها أخذ عينات على مستوى الأفراد داخل كُلّ وحدة من تلك الوحدات ذات مستوى أعلى. وتعد الأخطاء المعيارية بالنسبة إلى العينات متعددة المراحل، أكبر بكثير من الأخطاء المعيارية بالنسبة إلى العينات العشوائية البسيطة مع وجود عدد الترصدات أو الحالات (N) نفسها. وإن استخدام العينات العشوائية البسيطة، (SRS) - في وقت تتم فيه الدعوة إلى استخدام النوع الآخر دون مبرر - لا تظهر الحقيقة الكاملة للخطأ المعياري لكُلّ متنبِّئ على حدة، وبذلك تبرز نتائج إيجابية كاذبة (Thomas and Heck 2001).

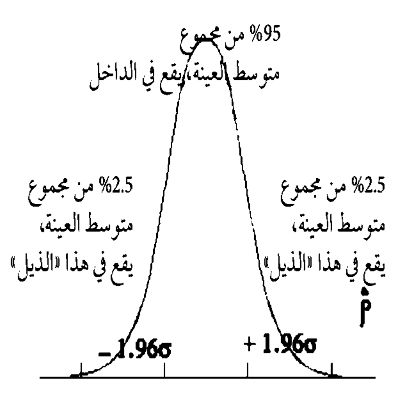
ويمكن استعمال مقاربات متعددة داخل البحث التقليدي لتكييف الأخطاء المعيارية بالنسبة إلى التصاميم المعقدة للعينة. وتُعرف معاملات التصحيح الأولى باسم (DEFF)، أي (تأثيرات التصميم)؛ كما تستخدم البرمجيات الأكثر حداثة، خطية تايلور (Taylor Linearization)، لتقدير الأخطاء المعيارية المصححة؛ وهذه علاجات فعالة على الرغم من عدم استخدامها من قبل كُلّ الباحثين.

ومع ذلك، يصبح اختبار الدلالة إشكالية أكثر، عندما يريد الباحثون تحليل بيانات ليست عينات عشوائية مستخلصة بشكل منتظم. ويصادف دارسون - بشكل متزايد - مجموعة بيانات مأخوذة من سجلات تنظيمية، أو مشتقة من معلومة مأخوذة من الشبكة العنكبوتية (The Web)، أو من مصادر أخرى كبيرة. وليست هذه المجموعة من البيانات مستمدة عشوائياً من الساكنين معروفة، على الرغم من احتمال أن تكون كبيرة جداً. والعبارة التقنية التي يمكن إطلاقها على هذا النوع من مجموعة البيانات (Dataset)، هي العينة المقبولة أو المريحة (Convenience Sample). أما بالنسبة إلى هذا النوع من العينات، فإن الاختبارات التقليدية للدلالة الإحصائية التي تفترض أن الباحث يحلل عينة عشوائية بسيطة مستمدة من ساكنين ما، هي اختيارات غير ملائمة تماماً.

(6) سنستخدم عبارة «أخذ العينات»، ومصطلح «المعاينة» بالتبادل لترجمة الكلمة الإنجليزية (Sampling)، دون أن يترتب عن ذلك تغيير في المعنى.

اختبارات تمهيدية وتبادلية

إنَّ للتنقيب في البيانات إجراءات متعددة، يمكنها تجنب مزالق مترابطة باختبار الدلالة في النموذج الأصلي التقليدي. وتشمل أحد حلول التنقيب في البيانات اختبار دقة الاستدلالات عبر المضاعفة والصلاحية التبادلية. وسنفصل القول في تلك الأفكار في قسم لاحق. ولكن يطبق حالياً - على نحو واسع - حلاً ثانياً، يعرف بـ العملية التمهيدية (Bootsrapping) على الاختبارات الدلالية، والنماذج الإحصائية التقليدية، وكذا داخل التنقيب في البيانات نفسها، وهي التقنية التي سنناقشها ابتداء (Mooney and Duval 1993).



الشكل رقم 1.2: توزيغ متوسط العينة.

وتستخدم المقاربة التقليدية لاختبار الدلالة (التي تسبق العملية التمهيدية) توزيع المعاينة (Sampling Distribution)، بغية تقدير الخطأ المعياري، ثم الدلالة الإحصائية أو قيمة - p لتقدير ما. إن توزيع المعاينة هي توزيع تم الحصول عليه نظرياً (ممثلاً في شكل صيغة رياضية أو في شكل رسم بياني على نحو مرئي، كما هو مبين في الشكل رقم 1.2)، الذي يصف كيفية علاقة التقديرات المستخلصة من عينات عشوائية مأخوذة من ساكنين ما، بالقيمة الحقيقية لذاك المعلم في الساكن.

لقد تم وضع افتراض (يسمى أيضاً افتراضاً معلمياً) من الافتراضات حول صلاحية توزيع معاينة نظرية، لدى استعمالها في تحليل معين بغية الحصول على فترات الثقة أو قيم-p لكُلّ انحدار خاص أو نموذج آخر. ولسوء الحظ، إذا ما تم

الإخلال بذلك الافتراض، فستكون الاستدلالات التي تم بلوغها حول الدلالة الإحصائية، خاطئة.

وتتجلى إحدى الطرق التي تلف هذه الصعوبة، في استخدام تقنية لا معلمية تعرف باسم العملية التمهيدية، إذ تَستعمِل استراتيجية تجريبية لتحديد الأخطاء المعيارية، والحصول على الدلالة الإحصائية أو قيم-p بالنسبة إلى مجموعة بيانات معينة أو تحليل تم إجراؤه، بدلاً من وضع افتراض حول شكل توزيع المعاينة.

إن العملية التمهيدية في شكلها المبسّط جداً، تَستعمِل العينة الوحيدة أو المنفردة لبيانات يحللها باحث ما، كما لو أنها تمثل السكّان برمتهم. وتستمد عملية التمهيد، عينات فرعية عشوائية عديدة من هذه العينة الوحيدة. وقبل استخلاصها الترصد الأول وضمه في عينة فرعية، تعيد استبدال تلك الحالة في العينة، ثم تختار عشوائياً ترصداً آخراً، مستبدلة تلك داخل التجمع (The Pool)، وتتكرر العملية إلى أن تنتج عينة ممهدة (Bootstrap Sample) مساوية من حيث العينة الأصلية، وهذا ما يعرف بالمعاينة باستبدال (Sampling with Replacement). وتتكرر هذه العملية لبناء - على ما يبدو - آلاف العينات الممهَّدة.

ولكل عينة من هذه العينات الممهَّدة العديدة، تُقدِّر - إذن - برمجيات إحصائية، ذات أهمية قد تكون متوسط عينة ما، أو معامل انحدار بالنسبة إلى متنبئ ما خاص ضمن نموذج معين. وستكون النتيجة، آلاف التقديرات المختلفة لتلك الإحصائية. ومن أصل ألف تقدير ممهد من هذه التقديرات، يتم بناء توزيع ما، يستخدم في تحديد الدلالة الإحصائية لأي إحصائية من العينة الأصلية (العينة اللا ممهدة) (Non Bootstrapped): بحيث يقيس واحد منها عدد التقديرات من أصل الآلاف منها التي تقع داخل مسافات متنوعة من مركز التوزيع، ويَحْسُب قيم -*p* المترابطة بتلك المسافات.

ولا تضع عملية التمهيد (أو تقنية من التقنيات ذات الصلة المسماة بالمطواة[7] (Jackknife) أي افتراضات حول شكل توزيع المعاينة. وكلاهما اجراءان تجريبيان محضان، يستعملان في قوة حوسبة قاسية (Brute Computing Power)، مُكَرِّرين

(7) أخذ هذا الاسم من «الخنجر السويسري»، لامكانية استعماله في أشياء متعددة بشكل مفيد جداً (المترجم).

انحداراً كاملاً أو تحليلاً آخراً، ربما ألف مرة. ونتيجة لذلك، تأخذ عملية التمهيد وقتاً إجرائياً معتبراً، ولو على الحواسيب السريعة.

ومع ذلك، يستخدم علماء التنقيب في البيانات أحياناً، منهجية قوة قاسية (Brute Force) أخرى لاختبار الدلالة - المعروفة باسم اختبار المبادلة (Permutation Test) أو اختبار دقيق (Exact Test)- في سياقات تستحيل فيها الدلالة التقليدية. لنعتبر أن نموذجك التنبؤي يتألف من خمسة متنبئات، ومتغير تابع واحد، Y. وكل خانة في جدول ممتد أو مصفوفة بيانات، تمثل إحدى متغيرات المتنبأ أو المتغير التابع. (تمثل السطور، والناس أو الحالات). إن برمجيات اختبار المبادلة، تخلط القيم داخل خانة ما. على سبيل المثال، إن القيم الموجودة داخل الخانة من أجل Y - المتغير التابع أو المستهدف - قد تبدَّل عشوائياً بقيم أخرى موجودة سلفاً في تلك الخانة، وتنتمي إلى حالات أو ترصدات أخرى. وهذا الخلط أو التبديل يخلط - عن قصد - قيم Y عبر الترصدات.

إن الإبدال (الخلط) يدمر أي بنية (أو علاقات) كانت موجودة سلفاً بين متنبئات X وY المبدلة حالياً. على سبيل المثال، قبل الإبدال قد يكون هناك ارتباط إيجابي بين X وY: الأفراد الذين كانت لهم قيمة عالية على X قد تكون لهم أيضاً قيمة عالية على Y، وفي الغالب، إن أولئك الذين كانت لهم قيماً منخفضة على X، لهم قيماً منخفضة على Y. ولكن من خلال القيام بخلط قيم داخل خانة Y، سيكون فرد ما ذو قيمة ما على X، مرتبطاً حالياً بقيمة شخص آخر على Y. لقد تمت إزالة البنية السابقة للارتباط (Correlation)، واستبدالها بالعشوائية. ولكن لاحظ أن القيمة المتوسطة للمتغير Y والانحراف المعياري لـ Y، سيتم الاحتفاظ بهما.

ومن ثم، يشغّل برنامج إحصائي ما، النموذج التنبؤي نفسه الذي شغّله سابقاً بالنسبة إلى البيانات الأصلية الحقيقية، وحالياً بالنسبة إلى هذه البيانات المجمّعة أو الممزوجة. وسيمنح ذلك مقياس تناسب - R^2 مثلًا - بالنسبة إلى مجموعة البيانات الممزوجة والمجمّعة. (كما يمكن التركيز على إحصائية أخرى، مثل التركيز على معامل ما لمتنبئ خاص. ويبقى المنطق نفسه ساري التطبيق).

وتتكرر هذه العملية من خلط نموذج ما وحسابه بعد ذلك، العديد من المرات،

قد تصل إلى ألف مرة. ومن ثم، فإن للباحث حالياً قيمة R^2 للبيانات الصحيحة والحقيقية، وكذا لألف قيمة أخرى من القيمة الإحصائية R^2 ، بواقع واحدة لكُلّ من العينات ذات بيانات عشوائية أو ممزوجة. ويأخذ ذلك وقتاً حاسوبياً كثيراً بشكل واضح. ويمكن لباحث ما أداء هذا الإجراء في نهاية مشروع تحليل ما، فقط عندما يكون واثقاً جداً بالنموذج النهائي، ويريد مستوى دلالي لذلك النموذج.

وبعد ذلك، يقارن الباحث قيمة R^2 (أو إحصائية أخرى ذات أهمية) في النموذج ذي البيانات الحقيقية بالقيم الموجودة في النموذج المختلط. لنتصور **أولاً** إمكانية أن يكون R^2 بالنسبة إلى النموذج الحقيقي، أكبر من القيم بالنسبة إلى ألف نموذج ممزوج برمته. وبعدئذٍ، يمكن لنا استنتاج أن احتمال الحصول على R^2 عَرَضاً هو أقل من واحد في الألف ($p \leq 0.001$)، على اعتبار أننا فحصنا 1000 عينة، ولا تملك أي عينة R^2 بهذا الحجم. إن ألف نموذج من نماذجنا التي تحتوي على بيانات ممزوجة، تجسيد واقعي للحظ: إذ من خلال تجميع البيانات، نكون قد قسنا فقط عدد المرات التي حدث فيها حجم معين لـ R^2 «بمحض الصدفة».

وبعدها، تصور أن من بين ألف نموذج عشوائي (مختلط)، عشرة نماذج لها R^2 مساوية (أو أكبر من) R^2 بالنسبة إلى النموذج ذي البيانات الحقيقية. ثم، إن احتمال الحصول - صدفة - على R^2 التي تم إيجادها بالنسبة إلى النموذج الحقيقي، هو 10/1000؛ أي إن $p = 0.01$. في المقابل، إذا أفضت عملية الخلط إلى 500 من أصل 1000 عملية بحيث يكون لها R^2 مساوٍ أو أكبر من R^2، المحصل عليها بالنسبة إلى عينة البيانات الحقيقية (غير المختلطة) (Non-Shuffled)، فإن الدلالة الإحصائية بالنسبة إلى بيانات حقيقية لشخص ما، هي 0,5 وسيكون من السهل وقوع النموذج الحقيقي الذي لا دلالة له إحصائياً عند مستوى 0.05، فقط بمحض الصدفة.

ويعد هذا الإجراء التبادلي، نوع من اختبار دقيق (Exact Test)، لا يفترض افتراضات حول شكل توزيع R^2 ، أو أي إحصائية أخرى خضعت للفحص. كما يعد هذا الإجراء أيضاً شكلاً من أشكال محاكاة مونتي كارلو (Monte Carlo Simulation).

• • •

ولتخليص هذا القسم حول اختبار الفرضيات، نزعم أن النقاط الرئيسة الواجب تذكرها هو أنَّ اختبار الدلالة، يقوم بدور حاسم في المنهجية الإحصائية التقليدية، حيث تستعمل للبتّ في المعاملات أو التأثيرات التي من المرجح أن تختلف عن الصفر في مجموع السكّان الكبير الذي أخذت منها العينة. ومع ذلك، اشتكى النقاد من أن ممارسات الباحثين اليومية في النمذجة التقليدية، غالباً ما تسيء استخدام اختبار الدلالة، مخلفة أخطاء معيارية صغيرة بشكل غير سليم، ونتائج إيجابية كاذبة عديدة. وإنّ أكثر مشاكل اختبار الدلالة خطورة، تحدث عندما يضيف واضعو النماذج متنبئات عديدة إلى النماذج خصوصاً لدى بحثهم في مئات المتنبئات قبل بتّهم في المتنبئ الذي يُضَمّ في نموذج ما.

وفي ردّ فعل على هذه الأخطاء، قال بعض المختصين في التنقيب في البيانات، وبعض المتنبئين، بالتخلي عن اختبار الدلالة جملةً وتفصيلاً (Armstrong 2007). إن معظم المختصين في التنقيب في البيانات، ليسوا بتلك الشدة، ولم يرفض معظمهم اختبار الدلالة برمته؛ وإنما ركزوا بشكل أكبر على المضاعفة والصلاحية المتبادلة باعتبارهما بديلين عن اختبار الدلالة عند تقييم نموذج تنبؤي.

علاوة على ذلك، إلى حدود منح تطبيقات التنقيب في البيانات، اختبارات الدلالة للمتنبئين الفرديين، فهي تستخدم - على الأرجح - اختبارات الدلالة التي تقوم سواء على عملية التمهيد، أو على اختبارات المبادلة، مما يسمح بتجنب العديد من المزالق المترابطة بالمقاربة التقليدية.

البرمجة اللا خطية في نماذج التنبؤ التقليدية

في نموذج انحدار عادي، تستخدم عدة متغيرات مستقلة أو متنبئات (لندعوها X_1، وX_2، وX_3) للتنبؤ بمتغير تابع (لندعوها Y). وقبل إنتاج نموذج ما، عادة ما تتحول تلك المتنبئات التي تمثل الفئات الاسمية (مثل الجمهوري، والديمقراطي، والمستقل)، إلى مجموعة من متغيرات وهمية أو صورية (Dummy Variables)، بحيث يأخذ كُلّ منها قيمة صفر أو واحد.

وأما المتغيرات المستقلة المستمرة (Continuous Independent Variables)،

مثل العمر، أو الدخل، أو سنوات التعليم، المسماة بقياسات الفاصل الزمني (Interval) أو قياسات النسبة (Ratio)، فهي عادة ما تدخل ضمن انحدار في شكل بسيط، مثلاً، العمر بالسنوات، أو الدخل بآلاف الدولارات، أو التعليم بالسنوات. إن معامل الانحدار بالنسبة إلى تلك المتغيرات الأخيرة قد يُفسَّر تقليدياً باعتباره التغيير في Y، المرتبط بزيادة وحدة واحدة (One-Unit) في X، فيما يتم التحكم في المتغيرات المستقلة الأخرى.

ويفترض هذا التفسير أن العلاقة بين X ما وY، علاقة خطية: أي إن زيادة وحدة واحدة من X في النهاية السفلى من سلم X، مرتبط بالقدر نفسه من التغيير في Y باعتباره زيادة وحدة واحدة لـ X في القيم العالية لـ X. وبتعبير آخر، إن رسم X مقابل Y على الرسم البياني قد ينتج خطاً مستقيماً. ولكن ماذا لو اقتحمنا الشكّ في امكانية أن تتنوع العلاقة عبر قيم مختلفة لـ X ؟ (سينتج ذلك منحنية (Curve) من المنحنيات أو خطاً ملتوياً (Wiggly Line)، إذا ما تم رسم X مقابل Y).

في بعض الحالات، يكون من البساطة نسبياً استبدال فترة زمنية مستمرة أو متنبئ نسبة بمجموعة من المتغيرات الوهمية أو الصورية (بحيث يأخذ كُلّ منها قيمة واحد أو صفر) التي ستمكننا من إدراك إمكانية وجود علاقة لا خطية بين X ما، أو Y. ومثال ذلك، عندما تستخدم أعوام من التعليم بصفتها متنبِّئاً. يفترض العديد من الباحثين أن تأثير أعوام من التعليم غير خطي، ومن ثم إعادة ترميز التعليم في مجموعة متغيرات وهمية - على سبيل المثال أقل من خريج مدرسة ثانوية، خريج مدرسة ثانوية، وكلية ما، ودرجة الباكالوريوس، درجة الماجستير أو درجة أعلى. وبعدها، تستطيع تلك المتغيرات - وهي تدخل بصفتها مجموعة ضمن تحليل انحدار تقليدي بفئة محذوفة، تعمل عمل فئة مرجعية - اجتلاب علاقات خطية بين التعليم وY على مستويات مختلفة من التعليم.

وحسب العديد من المتنبئين الآخرين، مع ذلك، لا يدرك باحث ما في وقت مبكر ما إن كانت العلاقة بين متغير X خاص وY، علاقة خطية. وتفترض الممارسة المعتادة في بحث كمي تقليدي، العلاقة الخطية، اللهم إلا إذا كان لشخص ما داع قوي لتبني عكس ذلك الطرح. ومن ناحية، يعد ذلك مجرد مسألة وقت وجهد:

وتستغرق مسألة فحص اللا خطية بالنسبة إلى العديد من المتنبئات، وقتاً طويلاً للغاية.

ومع ذلك، يحدث شيء آخر أكثر أساسية من مجرد الوقت والراحة، ذلك بأن قسماً كبيراً من الإحصائيات التقليدية قام على تصور الارتباط - أي المدى الذي يحقق فيه متغير ما زيادة في القيمة، ويخضع الآخر أيضاً لتغييرات. ويمكن لمجموعة بيانات برمتها أن تُمثَّل بمصفوفة الارتباط (Correlation Matrix)، أو مصفوفة تباين - التغاير (Variance-Covariance Matrix) التي تلخص العلاقات بين المتغيرات.

ولسوء الحظ، إن معامل الارتباط (Correlation Coefficient)، يقيس فقط العلاقة الخطية بين أي زوج من المتغيرات، وتهمل أي مظهر لا خطي. وتبسط - أحياناً - مسائل، فتفرض حلّاً غير مناسب. ولكن مع تطور التعلم الآلي وطرق أخرى كثيفة حاسوبياً، لم يبق هذا التبسيط ضرورياً. وتتوفر طرق «آلية» جديدة يمكنها البحث عن علاقات لا خطية، والعمل على صياغتها. وفي بعض الحالات، ستُنتِج هذه الطرق تنبؤاً أكثر دقة.

إنَّ هذه الأدوات الجديدة من أدوات التنقيب في البيانات الأكثر بساطة، تمكن الباحثين فقط من تصور بياناتهم: لرؤية العلاقات اللا خطية بين متغيرين أو أكثر باعتبارها صوراً، انطلاقاً من رسوم بيانية بسيطة أو مخططات التشتت (Scatterplots) إلى تصورات أكثر خيالاً، تمثل أسطحاً منحنية يمكن إدارتها، والنظر إليها من زوايا عديدة. ففي غامب (JMP)، مثلاً، تدعى إحدى أدوات التصور الأكثر إفادة، المحلل أو المِرسام (Profiler). وبعد تشكيل نموذج ما، يمكن للمرء استعمال هذه الأداة لمعرفة مدى تأثر قيم أي متغير كان، في وقت تتغير فيه قيم متغيرات أخرى.

وبعيداً عن التصور، يمكن لإجراءات التنقيب في البيانات الأخرى، توليد نقاط التوقف (Breakpoints)، بشكل آلي بالنسبة إلى متغيرات مستقلة مستمرة بهدف اجتلاب تأثير لا خطي لـ X ما في Y. مثلاً، قد يشير تحليل شجرة انحدار (CART) ما (التصنيف وشجرة الانحدار - نوع من تقسيم بيانات أو نموذج شجرة) إلى تأثيرات

لا خطية للدخل في Y، لتجد نقاط توقف دخل هام، بواقع 20,000$ و60,000$ و90,000$، و 150,000$.

وتشمل طريقة بديلة لاستكشاف علاقات لا خطية في البيانات، عملية يطلق عليها مختصون في التنقيب في البيانات اسم توزيع الخانات (Binning). وعموماً، يشمل توزيع الخانات، تحويل متغير رقمي مستمراً مثل الدخل داخل مجموعة من الفئات أو خانات منظّمَة. ولهذا، فعوض تمثيل الدخل بالدولارات التي تتراوح ما بين الصفر و1,000,000$ وأكثر، يصنف توزيع الخانات حالات أو أشخاص إلى فئات مثل صفر إلى 5000$؛ 5001$ إلى 15,000$؛ و15001$ إلى 25,000$؛ وهكذا. وثمة مصطلح آخر يستخدم في هذا الصدد، يدعى التفريد (Discretization): إذ يجعل من فئات منفردة شيئاً كان مستمراً.

وثمة نوع مفيد - بشكل خاص - من توزيع الخانات، يدعى توزيع الخانات الأمثل (Optimal Binning) أو التفريد القائم على الأنتروبي (Entropy-Based Discretization) للتعامل مع العلاقات اللا خطية. إنها تموضع نقاط التوقفات بين الخانات على نحو يعظم تنبؤ متغير تابع Y. وبتعبير آخر، تختار الحدود المخصصة لكُلّ خانة من قبل البرمجيات على نحو يجعل الحالات في كُلّ خانة مختلفة قدر الإمكان عن خانات أخرى من حيث قيمها على Y (Witen, Eibe, and Hall 2011, 316). وهذا مفيد جداً في تحديد علاقات لا خطية بين منتبئ ومتغير تابع.

وسنقدّم أمثلة في أقسام لاحقة، ولكن في هذه المرحلة، إن الفكرة الرئيسة الواجب تذكرها، هي أنه عند استخدام طرقاً إحصائية تقليدية، سيستغرق منا ذلك وقتاً طويلًا، وأحياناً تكون مسألة حظ لتحديد علاقات لا خطية بين كُلّ متغير من المتغيرات المستمرة المستقلة العديدة وبين المتغير التابع، وأنه نتيجة لذلك كله، يكون اعتيادياً أو من الشائع التعامل مع العلاقات باعتبارها خطية. ويقدم التنقيب في البيانات حالياً، أدوات متعددة لأتمتة (Automate) البحث في علاقات لا خطية، وما ذلك إلا سبب دفع - جزئياً - نماذج التنقيب في البيانات للميل إلى التنبؤ - على نحو أكثر دقة - بنماذج مشابهة لانحدار تقليدي.

تفاعلات إحصائية في نماذج تقليدية

غالباً ما تبحث النمذجة التقليدية عن تقدير المساهمات المتصاحبة لعدة متغيرات مشاركة (Covariates) للتنبؤ بمتغير النتيجة (Outcome Variable)، Y. على سبيل المثال، قد يُنَظّر باحث في التعليم حول دور الإعداد الأكاديمي لطالب ما في المرحلة الثانوية، ووضع العائلة الاجتماعي الاقتصادي، ومتطلبات الشغل، والدعم المادي في نموذج ما ليتنبأ بالطلاب الذين يتركون الدراسة. قد يكون هدف الباحث، تحديد أهم متنبئ أو أكثر تأثيراً في خطر ترك الدراسة في الكلية (انظر مثلاً Attewell, Heil, and Reisel 2011).

ولكن، يرى تشارلز راجين (Charles Ragin) (2008) أن العديد من المشاكل الاجتماعية، تحتاج إلى منطق مختلف تماماً عن منطق الهدف المذكور، وذو هدف مختلف: ويتجلى في فهم تراكيب عوامل مترابطة بنتائج مختلفة، عوض إبعاد دور المتنبِّئات الفردية. وعبّر راجين عن ذلك بـ«تضبيط أو تهيئة الشروط مقابل المتغيرات المستقلة». ويمكن دمج «تضبيط الشروط» في انحدار تقليدي ونماذج مماثلة من خلال ضم بنود التفاعل بين المتنبئين (انظر لمزيد من التفصيلـ Aiken and West 1991 and Jaccard and Turrisi 2003).

وإذا كان من الممكن ضم بنود التفاعل في نماذج تقليدية، فلا يعني ذلك أن يقوم الباحثون بذلك بصورة روتينية؛ على العكس من ذلك، يشتكي كُلّ من إلويرت ووينشيب (Elwert and Winship 2010) من أن الأغلبية الساحقة من الدراسات الكمية المنشورة في علم الاجتماع، تنقل فقط التأثيرات الأساسية (نموذج ذو متنبئات متعددة، دون تفاعلات). ومن ناحية، تحدث إزالة التفاعلات هذه، من النماذج التنبؤية، لأن (بحسب هذين المؤلفين) العديد من الباحثين أساؤوا فهم معنى معامِلات التأثير الرئيس في الوقت الذي تتم فيه «التحكم في» متغيرات مشاركة أخرى. بالإضافة إلى ذلك، يصدر غياب تأثيرات التفاعل في مقالات بحثية منشورة عن مشاكل عملية: ثمة أعداد هائلة من تفاعلات محتملة بين المتنبئين. ومن أصل 8 متنبئات، توجد 28 تفاعلات في اتّجاهين (تركيبات)، زائد تفاعلات إضافية ذات ترتيب أعلى. وكيف يحدد باحث ما التفاعلات الاستتباعية (Consequential) من

بين ثمانية وعشرين تفاعل محتمل؟ ستستغرق عملية بناء كُلّ متغير تفاعل على إلويرت (Elwert) ووينشيب (Winship)، أن معظم صانعي النماذج من الاجتماعيين، يهملون هذه المهمة ويقتصرون على نماذج التأثير الأساسية.

ويفيد مضمون انتقادات كُلّ من راجين وإلويرت ووينشيب بأنَّ الاعتماد على تأثيرات رئيسة في نماذج إحصائية تقليدية، نقطة ضعف خطيرة، وأن على الباحثين التركيز أكثر على تحديد تفاعلات معقدة بين المتنبئات المتعددة.

وقد توصلت إلى ذلك تقنيتان في التنقيب في البيانات بشكل سريع - سيتم تفصيل القول فيهما في فصول لاحقة (التصنيف وشجرة الانحدار CART)، وفي مربع للكشف عن التفاعل التلقائي (CHAID) - من خلال اختبار آلاف التفاعلات الممكنة أو التركيبات من بين المتنبئين، لتحديد التفاعلات المستتبعة لمتغير تابع خاص والتفاعلات غير المستتبعة. وبمجرد تحديدها، تستخدم بعد ذلك تلك التركيبات من القيم أو التفاعلات للتنبؤ بمتغير تابع أو هدف ما بطريقة منسجمة مع توصية راجين لدراسة «تهييئات الشروط»، عوض «متغيرات مستقلة».

كما بلغت تقنيات أخرى من التنقيب في البيانات تأثيراً مماثلاً، من خلال توليد ما يعادل التفاعلات بشكل آلي، داخل نماذجها التنبؤية. وتعد نماذج الشبكة العصبية مثال من الأمثلة التي سيتم وصفها لاحقاً.

وفي الغالب، يمكن لنماذج التنقيب في البيانات، التفوق على نماذج إحصائية تقليدية من حيث التنبؤ، أو في نسبة التباين التي تم شرحها، ذلك بأن نماذج كثيرة جداً، تهمل التفاعلات بين المتنبئات (سواء من خلال إزالتها بأكملها أو ضم فقط قليل منها، من أصل تفاعلات محتملة عديدة)، في حين تعد طرق التنقيب في البيانات أكثر شمولية أو دقة في تقييمها للتفاعلات واستعمالها من بين التنبؤات.

الجدول رقم 1.2: مفارقات بين النمذجة التقليدية والتنقيب في البيانات.

القضية	المنهجية التقليدية	التنقيب في البيانات
زيادة القوة التنبؤية	- ليس التنبؤ محط تركيز رئيسي - انخفاض التسامح مع قياس R^2	- التنبؤ محط تركيز رئيسي - تُقيَّم قوة تنبؤية عالية
اختبار الدلالة	- أساس التعميم - الحسم في تقييم الفرضيات وتفسير الآليات - بعض ممارسات الاختبار تعيبها التعددية	- تعميم من لدن الصلاحية المتبادلة بدل اختبار الدلالة - بعض التقنيات، «علب سوداء» (غياب أي معلمات مفيدة)
المعاينة	- اختبار الدلالة مرتبط بافتراضات المعاينة - كل العينات يتوقع أن تكون عينات عشوائية بسيطة أو عشوائية معقدة.	- انتشار تقنيات البوتسرابينغ أو العملية التمهيدية واللا معلمية - إقرار العينات المقبولة
علاقات لا خطية بين X وY	- غالباً ما يتم تجاهلها أو إهمالها	- تعريف آلي جزئياً
تفاعلات بين المتنبئين	- غالباً ما يتم تجاهلها أو إهمالها - النزوع المسبق للتأثيرات الأساسية	- تعريف آلي جزئياً للتفاعلات والتأثيرات غير المتجانسة

يلخص الجدول رقم 1.2، المفارقات المختلفة التي استخلصناها بين النمذجة الإحصائية التقليدية والتنقيب في البيانات. وفي الفصول الموالية، سنشرح كيف

تؤدي طرق التنقيب في البيانات أداءً مختلفاً، بل يمكن القول إنها أفضل من المنهجية التقليدية، ومن ثمة، فهي متفوقة في التنبؤ.

الاستنتاج

في هذا الفصل، قمنا بوصف جوانب متعددة من المقارنات بين منهجيات التنقيب في البيانات للتحليل الكمي من ناحية، وبين نمذجة إحصائية تقليدية من ناحية أخرى. كما ركزنا على الطرق التي يستتبع فيها منظور التنقيب في البيانات بعض الانتقادات لأكثر المنهجيات رسوخاً في تحليل البيانات. ماذا يعني هذا بالنسبة إلى العلاقة المستقبلية بين البحث في التنقيب في البيانات، وبين البحث الإحصائي التقليدي؟ من وجهة نظرنا، من المستبعد جداً أن يحل التنقيب في البيانات محل المنهجيات الإحصائية التقليدية. ومن المحتمل أن تكون عملية من عملية التهجين أكثر تطوراً، حيث يستخدم محللو المنهج الكمي - بشكل متزايد - بعض أدوات التنقيب في البيانات في عملهم، وحيث تشق بعض وجهات النظر الأكثر عمومية، الناشئة عن التنقيب في البيانات، طريقها صوب التنفيذ ونقل التحليل الكمي في العلوم الاجتماعية والسلوكية. ونتوقع أنواع التحولات القصيرة المدى التالية:

- سيولي الباحثون - بشكل متزايد - اهتماماً بإمكانية علاقات لا خطية بين المتنبِّئات والنتائج، من خلال الاستفادة من أدوات التنقيب في البيانات مثل توزيع الخانات الأمثل والأشجار لإنتاج متنبئات جديدة، تمثل النظم اللا خطية بشكل أفضل. وستضاف هذه المتنبئات المعدلة إلى نماذج معينة، وستساعد - في بعض الحالات - على الرفع من دقة النماذج التنبؤية. وسنقدم أمثلة على ذلك ضمن الفصول القادمة.

- سيصبح البحث عن التفاعلات الإحصائية بين المتنبئات، أكثر انتظاماً أو شمولية، وذلك بالاعتماد على أبحاث مُؤَتْمَتَة في التفاعلات الإحصائية المشابهة للنوع الذي سبق تقديمه في أدوات منمذجة غامب برو (JMP Pro)، و/ أو باستخدام أشجار القرار أو طرق التقسيم - مثل مربع لكشف عن التفاعل التلقائي (CHAID)، وشجرة الانحدار (CART) - التي تحدد التفاعلات. ومن ثم، فسيصبح - حسبما نتوقع - من الشائع رصد عملية ضم

العديد من بنود التفاعل في نماذج تنبؤية تقليدية، ستحسن من جديد دقة النموذج، R^2 ، أو التناسب (Fit).

- ومن الأرجح أن يفحص الباحثون بياناتهم لرصد تأثير عدم التجانس، وإمكانية اختلاف معاملات المتنبئات في نموذجها التنبؤي بصورة ملحوظة بالنسبة إلى مجموعات فرعية مختلفة داخل العينة أو الساكنة. إنَّ طرق التجميع التي يقدمها التنقيب في البيانات، والأدوات البارزة أدناه، مثل نماذج مختلطة (Mixture Models)، وانحدار الفئة الكامنة، تيسر البحث في التأثيرات غير المتجانسة وتميل إلى تقديم نظرة أكثر تعقيداً أو دقة لعمليات اجتماعية وسببية، مبتعدة عن نظرة «مقاس واحد يناسب الجميع».

- وقد يشهد الباحثون في النهج الكمي تحولاً في الرؤى - بحسب مدى تأثرهم بالتنقيب في البيانات في القادم من الأعوام - فيصرفون النظر عن هدف بناء نموذج تنبؤي واحد، الذي يعد جهدهم الأفضل، ويتبنون مقاربة مستلة من التنقيب في البيانات، تشكل نماذج تنبؤية متعددة مختلفة، مستخدمين في الغالب طرقاً متباينة للغاية، وتمزج بشكل مثالي التنبؤات من هذه الطرق المتعددة لإنتاج تنبؤ نهائي، أكثر دقة من ذلك المحصل عليه من أي نموذج كان. وتنجز هذا أدوات التنقيب في البيانات المعروفة باسم التعزيز (Boosting)، وطرق طقم منسجم الأجواء (Ensemble Methods) التي سنناقشها في فصول لاحقة. ويعمل هذا على تحسين الدقة التنبؤية المرتبط بالممارسات التقليدية.

- إننا نتوقع هنا بقاء اختبار الدلالة الإحصائية في العلوم الاجتماعية والسلوكية، وكذا في البحث التربوي والطب الإحيائي، على الرغم من الجهود المرحلية لإقناع المحررين لإلغائه لصالح التركيز على أحجام تأثير. ومع ذلك، نتوقع أن تؤثر ممارسات التنقيب في البيانات - بشكل متزايد - في الممارسات التقليدية الراهنة وتعديلها فيما يخص حساب مستويات الدلالة أو قيم -p ونقلها. ويمكننا سلفاً معرفة أن شعبية تقنيات إعادة المعاينة (Resampling)، مثل اختبارات تمهيدية وتبادلية، في تنام مستمر، ومرد

ذلك جزئياً إلى كون البرمجيات الحديثة والحواسيب الفائقة السرعة تيسر أكثر من عملية الحساب، ولأن هذه الطرق اللا معلمية لحساب الأخطاء المعيارية للتقديرات، لا تقوم على وضع افتراضات إحصائية غير قابلة للتصديق. كما تميل هذه الطرق الأكثر حداثة لحساب القيم-p لأن تكون أكثر تحفظاً من المقاربات القديمة - لتنتج في الغالب، أخطاء معيارية أكبر، ومن ثم إنتاج معاملات دلالية أقل. من المرجح - على ما يبدو - أن تقلص هذه الطرق مقدار الخطأ من نوع I، وسيُبدأ ذلك، عملية تقليص مقدار البحث غير القابل للإنتاج أو التكرار.

وقد تحدث خطوة أكبر نحو بلوغ هذا الهدف، إذا بدأ محررو المجلات يفرضون مقالات بحثية كمية لاستخدام طرق الصلاحية المتبادلة المألوفة في التنقيب في البيانات. وقدمنا باختصار منطق الصلاحية المتبادلة أعلاه، وسنقدم أمثلة ضمن الفصول المتتالية، ولكن تتمحور الفكرة الجوهرية حول كون كُلّ دراسة ستقسِّم بياناتها عشوائياً، وتختبر ما إن كان في إمكان نموذج تنبؤي ما، المتطور انطلاقاً من قسم من البيانات، التنبؤ بدقة، مستخدماً مجموعة من الترصدات التي لم تستخدم في إنتاج النموذج التنبؤي. وتعد الصلاحية المتبادلة شكلاً من أشكال المضاعفة التي «رفع الحاجز» (Raise the Bar)، لتقييم دليل تجريبي. وفي رأينا، سيكون لتبني الصلاحية المتبادلة، تأثير مهم ومفيد في العلوم الاجتماعية الكمية.

الفصل الثالث
استراتيجيات عامة مستخدمة في التنقيب في البيانات
صلاحية متبادلة

إنَّ تجريف البيانات - الذي يُعنى بالبحث في البيانات إلى غاية إيجاد علاقات ذات دلالة إحصائياً - عملية تستهجنها الكتب المدرسية التقليدية في الطرق، التي تعلّم الطلبة توليد فرضيات قبل بدء التحاليل الإحصائية. وقد نهضت مقاربة التنقيب في البيانات بتجريف البيانات إلى آفاق جديدة - ولكن ما يحسب لها، عدم مجاراتها المثل السيء للنموذج الأصلي التقليدي، فيما يخص اختبار الدلالة لما تكون هناك متنبّئات متعددة. إنها تركز - في المقابل - على طريقة بديلة من طرق تجنب نتائج إيجابية - كاذبة أو تجنب النوع الأول من الخطأ (Type I Error): أي إنها تركز على المضاعفة (Replication) عوض اختبار الدلالة، عبر إجراء ما يعرف بالصلاحية المتبادلة.

وقبل البداية لتحليل ما، متضمن للصلاحية المتبادلة، تفصل برمجيات التنقيب في البيانات الحالات داخل مجموعة بيانات ضمن مجموعات مختلفة، بحيث تُعهد كُلّ حالة أو ترصد، لمجموعة أو أخرى. (إن التخصيص العشوائي هنا أمر حاسم). وعادة ما تسمح برمجيات التنقيب في البيانات المستخدِم، باختيار نسبة الحالات من مجموعة البيانات الأصلية المخصصة لكُلّ مجموعة.

- تعرف مجموعة أو مجموعة فرعية عشوائية من الحالات أو الترصدات بعينة

التدريب أو عينة التقدير. وهذه هي مجموعة الحالات التي ستُحَلَّل أولاً، لإنتاج نموذج تنبؤي.

- ويتم إنتاج بعض طرق التنقيب في البيانات، وليس جميعها، المعروفة بعينة الموالفة (Tuning Sample) (وتدعى أحياناً عينة التثبت (Validation Sample). إنها تستعمل لتقدير بعض مَعْلَمات النمذجة التي تنتج تنبؤاً أمثل. على سبيل المثال، تمزج بعض تقنيات التنقيب في البيانات نماذج تنبؤية منفصلة ضمن أفضل مجهود نهائي في التنبؤ، مما يستدعي اتّخاذ قرار حول الكيفية التي يتم بها وزن التنبؤ، انطلاقاً من كُلّ نموذج من هذه النماذج لدى مزجها. وفي هذا السياق، يمكن استخدام هذه العينة العشوائية الثانية من الحالات - بيانات عينة الموالفة - لحساب أوزان بديلة، حتى يكون في مقدور مخطط الترجيح النهائي إنتاج التنبؤ الأكثر دقة، (وهذا ما يعرف بالأمثلية (Optimization)). وفي سياقات أخرى من سياقات التنقيب في البيانات، تُستعمل عينة الموالفة في المقابل، للبت في عدد المتنبئات التي ينبغي أن تدخل ضمن نموذج ما.
- وتعد مجموعة ثالثة من الترصدات المنتقاة عشوائياً، محورية في الصلاحية المتبادلة، وهذه عينة الاختبار التي تدعى أحياناً العينة المستبعدة (Holdout Sample. ولا يستعمل اختبار العينة - بأي حال من الأحوال - خلال إنتاج النموذج التنبؤي، وإنما يحتفظ به منفصلاً بأكمله، على نحو متعمد (أي مقيداً).

وخلال الخطوة الأخيرة، ضمن تحليل من تحليلات التنقيب في البيانات، يُطبَّق نموذج تنبؤي تم توليده باستعمال البيانات في عينة التدريب (Training Sample) (وأحياناً تشمل أيضاً بيانات عينة الموالفة(Tuning Sample)، على بيانات عينة الاختبار الجديد). ويولّد النموذج قيماً متنبَّأة للهدف بالنسبة إلى حالات الاختبار الجديدة هذه، وتقارَن تلك القيم المتَنَبَّأة، بالقيم الحالية المرصودة للهدف في بيانات الاختبار. ويُحْسَبُ الإحصاء التطابقي بالنسبة إلى هذه العينة من عينات الاختبار، مع توثيق مدى دقة تنبؤ النموذج المقَدَّر سابقاً بالمجموعة الجديدة من الترصدات.

وتؤدي الصلاحية المتبادلة وظيفة في التنقيب في البيانات مماثلة لتلك المقدمة من قبل اختبار الدلالة في المنهجية التقليدية: إنها طريقة من طرق تقييم تعميم نتائج البحث. ويمكنك أيضاً، التفكير في الصلاحية المتبادلة باعتبارها نوعاً من أنواع ضبط الجودة بالنسبة إلى نماذج التنقيب في البيانات.

ويتجلى الفرق في المنهجيتين في تناول التعميم، في كون أنه في النموذج الأصلي التقليدي، تكشف اختبارات الدلالة الإحصائية عن إمكانية تعميم النتائج المحصّل عليها من عينة معينة على السكّان الذين أخذت منها العينة عشوائياً. علاوة على ذلك، يعد تقييم التعميم (Generalizability)، تقييماً نظرياً أو افتراضياً، بحيث لا يملك الباحث بيانات حقيقية للسكان بأكملها. في المقابل، يعد اختبار التعميم في مجال التنقيب في البيانات، اختباراً تجريبياً، بحيث يُطَبَّق نموذج من النماذج التي تم تطويرها وأداؤها بشكل جيد في التدريب أو في عينة التقدير، على عينة مختلفة من بيانات حقيقية (عينة اختبار)، وتخبرنا دقة المطابقة (Goodness of Fit) برأي الباحث في مدى تعميم النموذج على البيانات الجديدة. وفي حالة التنقيب في البيانات، لا يتم التعميم من عينة إلى ساكّن، وإنما من عينة عشوائية إلى عينة عشوائية أخرى (أي من التدريب إلى عينة الاختبار).

ثمة متغيرات عديدة للصلاحية المتبادلة، إذ تعرف أبسطها باسم الطريقة المستبعدة (Holdout Method)، ومناسبة بشكل مثالي لتحليل بيانات ضخمة ذات ترصدات متعددة. وتُقسِّم مجموعة بيانات ما بشكل عشوائي إلى عينتين فرعيتين أو ثلاثة (عينات التدريب، والموالفة، والاختبار)؛ فتُستبعد عينة الاختبار، ولا تستعمل في تدريب النموذج التنبؤي. وإذا كانت مجموعة البيانات الأصلية كبيرة جداً، فإن هذا التقسيم العشوائي للعينة الأصلية إلى قسمين أو ثلاثة أقسام، لا يؤدي إلى إشكالية فقدان القوة الإحصائية عند تقدير النماذج التنبؤية. ستُتْرك حالات كثيرة في عينات التدريب الفرعية. لاحظ أنه في الطريقة المستبعدة، يسند كُلّ ترصد على حدة عشوائياً، إما إلى عينات التدريب الفرعية، أو عينات الموالفة الفرعية، أو عينات الاختبار الفرعية. وعليه، فإن كُلّ عينة فرعية تضم حالات أو ترصدات منفصلة على نحو كامل.

ومع ذلك، فالصلاحية المتبادلة، لا تحتاج إلى مجموعة بيانات كبيرة، بل يشتغل نوع مختلف من أنواع الصلاحية المتبادلة، المعروفة بالصلاحية المتبادلة ذات الطية-ك (K-Fold)، على مجموعة بيانات صغيرة وكبيرة. ويبدأ الإجراء بخلق عدد مختار (k) من عينات فرعية عشوائية، بحيث يتم في الغالب اختيار 10. وتُسحب الحالات أو الترصدات عشوائياً من العينة الأصلية، فتُسند إلى كُلّ عينة فرعية، إلى أن يصير لواحد منها، عدد k المختار عشوائياً من العينات الفرعية في جميع الحالات. وتضم كُلّ حالة عدداً واحداً k من العينة الأصلية.

وستعمل إحدى تلك العينات الفرعية لـk بداية، باعتبارها مجموعة بيانات اختبار، في حين يتم تجميع العينات الفرعية الأخرى k-1 لتشكيل مجموعة تدريب. ويقَدَّرُ نموذج ما بالنسبة إلى مجموعة تدريب مجمّعة، وبعدها يتم اختبار هذا النموذج التنبؤي على العينة الفرعية ذات مجموعة اختبار واحد؛ فتنتج إحصائية تطابقية أو قياس خطأ.

وهكذا، يتكرر هذا الإجراء عدداً من المرات k في جميع الحالات، بحيث تقوم كُلّ عينة فرعية k بدور مجموعة بيانات الاختبار مرة واحدة فقط، في حين تمثل الطيات المتبقية المختلطة، بيانات التدريب. وإن إحصائية التطابق النهائية التي تنقلها البرمجيات هي معدل إحصائيات التطابق بالنسبة إلى عينات الاختبار عبر كُلّ عمليات k.

ومهما يكن شكل الصلاحية المتبادلة المنتقاة (وهناك متغيرات إضافية)، فإن النقطة الحاسمة التي ينبغي تذكرها، هو أنه عندما يتم تقييم الدقة التنبؤية لنموذج ما، يجب دائماً النظر إلى إحصائيات التطابق من أجل العينة المستبعدة أو عينة الاختبار. وتنقل بعض البرمجيات، إحصائيات التطابق بالنسبة إلى عينة التدريب أيضاً، ولكن تبقى إحصائية التطابق للعينة المستبعدة أو لعينة الاختبار، الإحصائية المهمة دائماً.

ولفهم سبب اعتماد مختصي التنقيب في البيانات فقط على إحصائيات التطابق بالنسبة إلى عينة الاختبار، ينبغي التحول إلى ظاهرة مهمة أخرى، تعرف باسم التدريب المفرط (Overfitting).

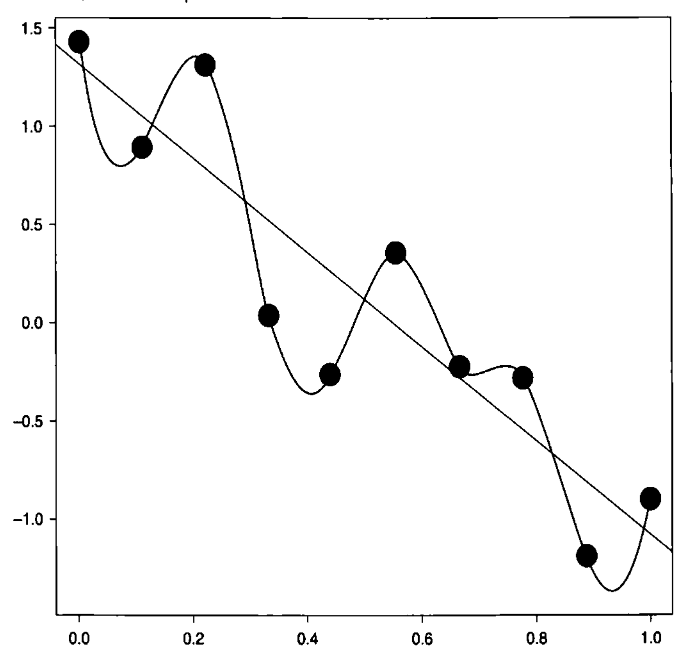
التدريب المفرط

إن للتنقيب في البيانات نقاط ضعف خاصة بها، وما التدريب المفرط إلا واحد منها. وتعد بعض تطبيقات التنقيب في البيانات ناجعة جداً في بناء نموذج تنبؤي،

بحيث تبني شيئاً معقداً جداً، سيعمم على عينات أخرى. وسيكون ذلك سهلاً جداً ومفصلاً بيانياً (انظر الشكل رقم 1.3).

يمكن لنموذج ما، تفسير العلاقة بين X وY في هذا المخطط من خلال مواءمة خط مستقيم، يمثل القيمة التنبؤية لـ Y لقيم متنوعة من قيم X. وتمثل المسافة العمودية من كُلّ نقطة بيانات إلى الخط المستقيم، خطأ التقدير بالنسبة إلى كُلّ نقطة بيانات في ذلك النموذج البسيط، أي الفرق بين القيمة المتنبَّأة لـ Y والقيمة المرصودة لـ Y بالنسبة إلى كُلّ قيمة من قيم X.

وقد يقلص نموذج من النماذج الأكثر تعقيداً للغاية مقدار خطأ التنبؤ. ويمثل الخط المتموج (Wavy Line) معادلة من قبيلـ... $Y=a+bX+cX_2+dX_3+eX_4+fX_5...$. وكما يمكنكم رصد ذلك في المخطط، إن هذا الخط الأكثر تعقيداً، يمر بشكل مستقيم عبر كُلّ نقاط البيانات، مما يدل ضمناً على انعدام وجود أي خطأ تنبؤي ما.



الشكل رقم 1.3: بيانات التدريب المفرط.

ما العيب في اختيار نموذج أكثر تعقيداً إذا كان ذلك يعمل على تقليص الخطأ،

وينتج تنبؤاً أقوى؟ وقد يحذر مختصو التنقيب في البيانات من أن بعض المسافة التي تفصل كُلّ نقطة بيانات الخط المستقيم، قد ترجع احتمالاً، إلى خطأ القياس، أي إلى الضجيج (Noise). وباستخدامنا نموذجاً معقداً بشكل كبير - مثل الخط المتموج - لمواءمة نقاط تلك البيانات بشكل دقيق، لا نعني فقط مواءمة الإشارة (Signal)، وإنما أيضاً مواءمة الضجيج. إن النموذج المتموج المعقد في اصطلاح التنقيب في البيانات، يعد بيانات تدريب مفرطة. والتدريب المفرط أمر غير مرغوب فيه، لأن ذلك يعني أن النموذج المعقد لن يعمل بشكل ممتاز ما إن طُبق على بيانات أخرى، مثل بيانات الاختبار. لقد فُصّل النموذج حسب بيانات التدريب المفعمة بالضجيج، ومن ثم لن توائم نتيجة ما، بيانات أخرى بشكل ممتاز.

كيف يتسنى للمرء معرفة ما إذا كان نموذج أو معادلة ما، ذات تدريب مفرط أم عكس ذلك؟ عندما يُطبق النموذج التنبؤي (عادة في شكل معادلة) المشتق من عينة تدريب معينة، على عينة اختبار منفصلة بشكل كامل، وتحتوي على ترصدات أو حالات مختلفة، آنذاك يمكن للمرء مقارنة القيم المتنبَّأة المحصل عليها انطلاقاً من النموذج، بالقيم المرصودة في مجموعة البيانات الجديدة، وتحديد مدى مواءمتها. وتقدم هذه الخطوة الثانية تقييماً جديراً بالثقة لمدى صلاحية النموذج التنبؤي لبيانات لم تستخدم من قبل.

و«سيتراجع» التدريب المفرط أو يخفق في المساعدة على تنبؤ الاختبار أو البيانات المستبعدة لأن جزء من النموذج الذي وصف أنماط الحظ في بيانات التدريب (القسم ذو التدريب المفرط)، سيخفق في تنبؤ أي شيء مفيد في مجموعة البيانات الثانية أو مجموعة بيانات الاختبار. وسيكون هناك ضجيج عشوائي في عينة الاختبار العشوائي أيضاً، ولكن إذا كان الضجيج عشوائياً، فمن الطبيعي أن يكون الضجيج نفسه كما هو الحال في مجموعة البيانات الأولى. ومن ثم، لن يكون لها النمط نفسه، بل لن يكون لها أي نمط من الأنماط في واقع الأمر.

عادة ما ستكون إحصائية تطابقية لنموذج ما، تم حسابه بالنسبة إلى عينة تدريب، أفضل من تطابق النموذج نفسه المطبق على عينة اختبار (ذلك بأن بيانات الاختبار لن تكون ذات تدريب مفرط). وإذا ما وجد فرق كبير في إحصائية التطابق بين عينة

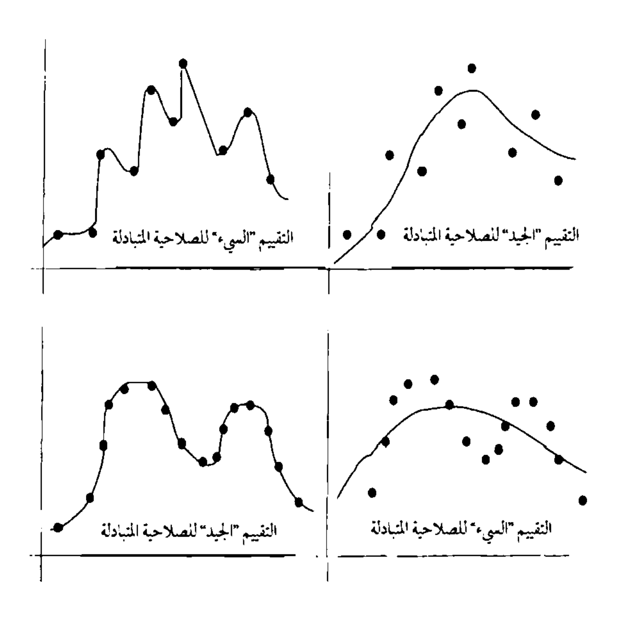
تدريب وعينة اختبار، فإن ذلك إشارة قوية على وجود تدريب مفرط في الحالة الأولى. وفي المقابل، إذا كان تطابق نموذج ما في عينة تدريب ما، وتطابق النموذج نفسه المطبّق على عينة اختبار ما، قريبين بما فيه الكفاية، فإن مؤدى ذلك انعدام وجود تدريب مفرط في الحالة الثانية: ومن ثم، فإن النموذج قادر على التعميم بشكل جيد.

وكي نختم، يبدو أن استخدام التنقيب في البيانات للصلاحية المتبادلة، منهجية أكثر صرامة لتجنب الخطأ من النوع الأول (نتائج إيجابية كاذبة) من استراتيجية اختبار الدلالة المألوفة في البحث الاجتماعي التقليدي. يقيم المرء دقة نموذج من نماذج التنقيب في البيانات من خلال علم الإحصاء التطابقي المحصل عليها لفائدة اختبار مختار بشكل عشوائي أو لعينة مستبعدة، وهذا يوفر قياساً جديراً بالثقة لتعميم النتائج.

ويقدم **الشكل رقم 2.3**، تصويراً مرئياً لإمكانية استخدام الصلاحية المتبادلة لتجنب التدريب المفرط. إن الربعين (Quadrants) الموجودتين في أعلى المخطط، هي تحاليل نقاط البيانات نفسها. وتظهر الربعية الموجودة في أعلى يسار المخطط، نموذجاً معقداً مطابقاً لهذه البيانات، أي منحى ذا نقاط انعطاف عديدة، حيث يطابق النموذج (الممثل بالخط) كُلّ نقاط البيانات إلى حدّ بعيد، مما سينتج تنبؤاً جيداً جداً بالنسبة إلى بيانات التدريب. ومع ذلك، قيل لنا بشأن الربعية، إن الصلاحية المتبادلة (CV) تخبرنا بأن هذا نموذجاً سيئاً جداً، لأن الإحصائيات التطابقية كانت تتقلص تدريجياً بشكل كبير عندما يتم تطبيقها على بيانات الاختبار. وكان النموذج الأصلي ذا تدريب مفرط بكل تأكيد.

وتشير الربيعتان أيضاً إلى مجموعة بيانات مستقلة، ولكنها مجموعة بيانات مختلفة انطلاقاً من النصف الأعلى للمخطط. وعلى الجانب الأيسر، تتم عملية مطابقة نموذج معقد من نماذج التنقيب في البيانات، ولكن قيل لنا إن إحصائيات التطابق للصلاحية المتبادلة لصالح هذا النموذج، هي تقريباً جيدة بالنسبة إلى عينة الاختبار، وعينة التدريب على حدّ سواء. وهكذا، يمكن أن نستخلص أن هذا نموذجاً قابلاً للتعميم، على الرغم من كونه معقداً؛ وليس نموذجاً ذا تدريب مفرط. وللتنبيه فقط، فإننا نحاول أيضاً أن نجرّب نموذجاً أكثر بساطة على البيانات نفسها. وتوضح ذلك الربيعية الموجودة في أسفل يمين المخطط. وبهذا النموذج الجديد، نجد أن

إحصائيات التطابق بالنسبة إلى عينة الاختبار، جيدة بالقدر نفسه بالنسبة إلى عينة التدريب.



الشكل رقم 2.3: الصلاحية المتبادلة

(**مأخوذة من**: www.cs.cmu.edu/~schneide/tut5/node42.html#figcvo).

يمكن استخلاص ثلاث دروس من هذه الرسوم التوضيحية:

- البساطة في نموذج ما، ليست دائماً جيدة (على الرغم من أننا عادة ما نفضلها)؛
- التعقيد في نموذج ما، ليس دائماً دليلاً على التدريب المفرط؛
- الصلاحية المتبادلة إجراء موضوعي، يمنع المرء من تقبل النماذج ذات التدريب المفرط.

التعزيز

لقد قلنا إن التنقيب في البيانات يؤكد أهمية التنبؤ الدقيق، وهو - منهجية مع النموذج الأصلي التقليدي - أقل تقبلاً لنماذج قادرة فقط على تفسير نسبة صغيرة من التباين في متغير مستقل. ولأن التنبؤ المعزز يُعدّ باعثاً قوياً بالنسبة إلى المختصين في التنقيب في البيانات، فإنهم طوروا تقنيات جديدة تعمل على تحسين التنبؤ. يبدو بعضها غريباً جداً عندما ينظر إليه من منظور نمذجة العلوم الاجتماعية التقليدية. ولكن هذه الاستراتيجيات - كما سنبين ذلك لاحقاً - غالباً ما تتفوق على النماذج التقليدية عندما يتعلق الأمر بالتنبؤ.

وما التعزيز (Boosting) إلا أحد هذه الاستراتيجيات، إذ تتعامل مع إنتاج النموذج باعتباره سلسلة من الخطوات. وقد يبدأ المرء مثلاً، بتقدير نموذج انحدار للتنبؤ بمتغير مستهدف مرصود Y. وإن تطابق النموذج ليس مثالياً، ومن ثم، فستكون لكُلّ ترصّد قيمة متبقية أو خطأ تنبؤ، أي الفرق بين القيمة المرصودة والمتنبّأة على Y بالنسبة إلى كُلّ حالة، أو $Y-\hat{Y}$.

وفي خطوة ثانية، يُقدَّر نموذج تنبؤي آخر باستخدام طريقة نمذجة مختلفة، ولكن هذه المرة من خلال تنبؤ القيم المتبقية (Residuals) انطلاقاً من النموذج الأول، عوض تنبؤ المتغير الأصلي التابع، Y. وينتج هذا النموذج الثاني أيضاً قيماً متنبأة، ولكن تظل بعض أخطاء التنبؤ قائمة. ولذلك، يمكن للقيم المتبقية من هذا النموذج الثاني - بدوره - تنبؤها بواسطة نموذج ثالث، وهكذا بالنسبة للعديد من عمليات تكرارية.

وأما الخطوة الأخيرة في تحليل معزز، فتتجلى في مزج معادلات التنبؤ المحصّل عليها من كُلّ خطوة (Ridgeway 1999). ويُنجز ذلك أحياناً من خلال توفير أوزان متناقصة لنماذج ناجحة، ومن ثم تجميع التنبؤات للحصول على تنبؤ وحيد أفضل لـ Y.

ويمكن للتعزيز أن ينتج تحسناً كبيراً في التطابق الأخير، أو الدقة التنبؤية لنموذج من نماذج التنقيب في البيانات، مقارنة بمقاربة تقليدية ذات الخطوة الوحيدة. وقد كتب ماتهياس شونلو (Matthias Schonlau) (2005) برنامج «الستاتا» (Stata)

(Program المعروف باسم «زيادة» (Boost) الذي يُطبِّق خوارزم تعزيز مألوف. وينقل أداؤه من خلال مثالين: انحدار خطي تقليدي وانحدار لوجيستي تدريجي؛ ففي السياق الأول، تنبأ نموذج انحدار المربعات الصغرى التقليدية العادية 21.3٪ من التباين (R^2)، في حين فسرت المتنبئات المتطابقة، والبيانات في انحدار معزز، 93.8٪ من التباين. أما بالنسبة إلى انحدار لوجيستي تدريجي، فقد صنف برنامج «الستاتا» التقليدي بشكل صحيح، 54.1٪ من الحالات بيانات الاختبار، ولكن التعزيز تنبأ بشكل صحيح بـ76.0٪ من الحالات في عينة من عينات الاختبار. وهذه زيادات ضخمة في القوة التنبؤية بفضل التعزيز.

وستَذْكرُ أن إحدى التفسيرات المألوفة المقدمة بشأن السبب وراء تفسير النماذج الإحصائية التقليدية (أي لا تعتمد التنقيب في البيانات)، التي تفسر في الغالب فقط نسبة صغيرة من التباين، تتجلى في حضور قياس الخطأ و/ أو في المفهوم الذي يفيد بأن بعض العوامل المهمة لم يتم قياسها، ومن ثم حذفت من النموذج. ومع ذلك، نرى هنا أن تقنية واحدة من تقنيات التنقيب في البيانات - التعزيز - يمكن أن ترفع نسبة التباين المفسرة بشكل كبير، مقارنة بنموذج تقليدي، مع استعمال - في الوقت نفسه - المتنبئات والبيانات نفسها بشكل دقيق باعتبارها النموذج التقليدي. وفي هذه الحالة، يعد الادّعاء بأنَّ خطأ القياس والمتغيرات المحذوفة هي المسؤولة عن تقليص التباين المفسر، ادّعاءً مجانباً للصواب.

وثمة شيء عن الأداء التنبؤي بشكل واضح لهذه النماذج التقليدية التي تعد أقل شأناً من منهجية التنقيب في البيانات. لقد كان التعزيز قادراً على إيجاد مزيد من البنية في البيانات، أكثر مما تستطيع المقاربة التقليدية القيام به. ولم يكن ذلك راجعاً إلى التدريب المفرط، لأن هذه الإحصائيات التطابقية المثيرة ليست موجهة للعينة العشوائية الأصلية لبيانات التدريب التي أنتجت النموذج التنبؤي، وإنما لعينة بيانات عشوائية منفصلة بشكل تام، بيانات الاختبار. لقد استخدم مقال شونلو بيانات مشكلة اصطناعياً. وقمنا بإنجاز تحليل مماثل لمعرفة ما إن كان أداء التعزيز جيداً أيضاً مع بيانات العالم الحقيقي. ونقل انحدار مربعات صغرى تقليدية عادية في الجدول رقم 1.3 أدناه، إذ يتم فيه تنبؤ لوغاريثم الدخل الشخصي من خلال متغيرات السوسيو الديموغرافية المتعددة، وذلك باستخدام بيانات مستقاة من مسح المجتمع الأميركي،

الذي أعده مكتب تعداد السكّان في الولايات المتحدة الأميركية (Census Bureau)، عام 2010. على الرغم من وجود عينة كبيرة، ومتنبئات عديدة، وجمع بيانات ذات جودة عالية تقنياً، فإن التباين المفسر كما يمثله انحدار R^2 هو فقط 29٪.

الجدول رقم 1.3: انحدار مربعات صغرى عادية تتنبأ لوغاريثم الدخل الشخصي.

مجموع الدخل الشخصي (log)	المعدل	SE	قيمة *p*
العمر	0.0674	0.0007	<.0001
العمر مربع (ممركز)	−0.0006	<0.0001	<.0001
أنثى	−0.5387	0.0021	<.0001
الهيبة المهنية	0.0211	<0.0001	<.0001
أسود	−0.0626	0.0037	<.0001
أميركي أصلي	−0.1804	0.0135	<.0001
آسيوي	−0.0489	0.0056	<.0001
عرق آخر (المرجع = أبيض)	−0.0162	0.0046	<.0001
أرمل	0.0707	0.0063	<0.0001
مطلق	−0.0522	0.0028	<.0001
منفصل	−0.1544	0.0061	<.0001
متزوج، الزوج غائب (المرجع = متزوج، زوج حاضر)	−0.1833	0.0070	<.0001
غير مواطن	−0.1609	0.0045	<.0001
مواطن مجنس (المرجع = مواطن بالولادة)	0.0213	0.0042	<.0001

أقل من المدرسة الثانوية	-0.2350	0.0041	<.0001
كلية ما، من دون درجة علمية	0.1019	0.0030	<.0001
درجة الزميلة	0.1528	0.0039	<.0001
درجة الباكالوريوس	0.2913	0.0035	<.0001
أكثر من درجة الباكالوريوس	0.4460	0.0042	<.0001
بريطانيا الجديدة	0.0971	0.0051	<.0001
منتصف المحيط الأطلسي	0.0951	0.0036	<.0001
وسط الشمال الشرقي	-0.0090	0.0034	<.0001
وسط الشمال الغربي	-0.0172	0.0047	<.0001
وسط الجنوبي الشرقي	-0.0827	0.0047	<.0001
وسط الجنوب الغربي	-0.0050	0.0038	0.1820
الجبال	-0.0460	0.0047	<.0001
المحيط الهادي (المرجع = المحيط الأطلسي الجنوبي)	0.0712	0.0036	<.0001

ملاحظة: ترصدات N = 1,226,925؛ ثابتة = 8.077؛ $R^2 = 0.2882$.

وفي الجدول رقم 2.3، يقارن هذا النموذج التقليدي بنماذج متعددة للتنقيب في البيانات التي استعملت البيانات ذاتها. إن السطر الأول يكرر R^2 بالنسبة لانحدار المربعات الصغرى التقليدية العادية أعلاه، في حين تنقل الأسطر الأخرى إحصائيات R^2 بالنسبة إلى أربعة نماذج مختلفة من نماذج التنقيب في البيانات، مستخدمة البيانات والمتغيرات المتطابقة. وفي كُلّ حالة، تفسر مقاربة التنقيب في البيانات - بشكل معتبر مزيداً من التباين أكثر من الانحدار التقليدي: إن لها قوة تنبؤية أفضل بكثير (على الرغم من أننا لم نشهد تحسنا كبير مقارنة مع مثال شونلو). وتستخدم هذه

النتائج بيانات حقيقية، لكن يتم عرضها هنا فقط من أجل أغراض توضيحية. وإذا ما كنا قد «أضفنا تعديلات نهائية» إلى نماذج التنقيب في البيانات بدرجة أكثر، من خلال تطويع معلمات متنوعة، ولأمكن لنا زيادة R^2 أبعد من ذلك.

الجدول رقم 2.3: أداء انحدار المربعات العادية الصغرى المعيارية، مقابل نماذج التنقيب في البيانات.

نوع النموذج	عينة اختبار R^2
المربعات العادية الصغرى	.288
شجرة التقسيم	.442
غابة نظام تمهيدي لتشغيل الحاسوب	.438
الشجرة المعززة	.436
الشبكة العصبية	.481

معايرة

المعايرة الاستراتيجية الأخرى من استراتيجيات التنقيب في البيانات لتحسين تنبؤ النموذج الذي انحرف أيضاً عن الممارسات التقليدية. وإن إحدى الافتراضات الإحصائية الكامنة وراء نمذجة الانحدار التقليدي هو كون - وعبر الطيف الترددي لقيم المتغير التابع Y - التقدير الأفضل لـ Y، يعد دائماً التنبؤ (الذي يُدعى $\hat{Y}$ أو Y- قبعة) المقدم من قبل معادلة الانحدار. ونتيجة لذلك، يجب أن يكون خط (Plot) القيم المتنبأة لـ Y مقابل قيم Y المرصودة، خطاً مستقيماً. وإذا كان الأمر كذلك، فسيعد النموذج، نموذجاً معايراً (Calibrated).

ولسوء الحظ، إن تحليلات بيانات العالم الحقيقي، سواء تعلق الأمر بخط ما أو

رسم بياني للعلاقة بين Y و $\hat{Y}$، هي في الغالب، علاقة خطّية عبر كثير من مجموع قيم Y، لكنها تنحرف عن خط مستقيم إما في قيم عالية أو منخفضة لـ Y أو فيهما معاً. ومن ثم، فإن الخطّ ينحني. وفي هذه الحالة، يُعد نموذج الانحدار نموذجاً غير معاير (Uncalibrated)، بحيث لا يتنبأ النموذج بدقة في القيم القصوى لـ Y كما تفعل في المدى المتوسط. وفي المقاربة التقليدية، يحاول باحث ما، تحديد متغيرات تنتج هذا النمط المنحني، وإضافة آخرين إلى نموذج الانحدار، آملين أن يتسبب ذلك في اختفاء الانحناء (Curvature).

يستخدم التنقيب في البيانات أحياناً، منهجية مختلفة. إذا كان نموذج ما غير معاير، كما أشير إلى ذلك بواسطة خط منحن لـ Y مقابل $\hat{Y}$، فإن الباحث قد يوائم نموذجاً متعدد الحدود(Polynomial) مع Y($Y= \hat{Y}+ \hat{Y}^2+ \hat{Y}^3+\ldots$)، أو دالة أخرى ناعمة مثل دالة الخُدة (Spline). ولا يضيف هذا الإجراء أي شيء إلى الفهم الموضوعي للعلاقة بين المتنبئات المتنوعة والمتغير التابع، لأن الباحث لم يكتشف سبب حضور المنحنى. ومع ذلك، حسّن هذا الإجراء دقة تنبؤ Y وطور مواءمة النموذج.

ويقدم الجدول رقم 3.3 توضيحاً لتأثيرات المعايرة في التباين المفسر، وذلك باستخدام نموذج انحدار المربعات الصغرى، وتنبؤ لوغاريثم الأرباح (Log of Earnings) بحيث تشمل المتنبئات: العمر، وتربيع العمر، والتحصيل العلمي (بمثابة مجموعة من متغيرات وهمية (Dummy Variables))، والمنطقة، والجنس، وساعات العمل، وأسابيع العمل. مرة أخرى، إن البيانات مأخوذة من مسح المجتمع الأميركي لعام 2010. وإن إضافة الحدود $\hat{Y}^2$، و$\hat{Y}^3$ و$\hat{Y}^4$ في معادلة الانحدار، يرفع من التباين المفسر من 0.52 إلى 0.59، مبيناً أن المعايرة يمكن أن تنتج تحسناً في الدقة التنبؤية.

إن التعزيز والمعايرة هما استراتيجيتان مألوفتان في التنقيب في البيانات، بحيث يوضح كلاهما التركيز القوي الذي يضعه التنقيب في البيانات على تحسين التنبؤ، وعلى الطريقة التي تخلّفُها في استراتيجيات تحليلية جديدة.

الجدول رقم 3.3: تأثير المعايرة في تناسب النموذج.

	R^2	خطأ جذر متوسط المربعات (RMSE)
نموذج انحدار المربعات العادية الصغرى الأساسي	.5237	2.28
المذكور أعلاه + حدّ تربيعي: $\hat{Y}^2$	.5929	2.11
المذكور أعلاه + حدّ تكعيبي: $\hat{Y}^3$	.5939	2.11
المذكور أعلاه + حدّ رباعي: $\hat{Y}^4$	.5949	2.11

تناسب القياس: مصفوفة الارتباك ومنحنيات جهاز

يستخدم مختصو علماء التنقيب في البيانات مصطلح تناسب (fit)، للإشارة إلى دقة نموذج تنبؤي، وتحديداً إلى مدى قرب قيم تنبؤية لمتغير هدف أو متغير تابع من قيم مرصودة لذلك المتغير. وإن القياس الأبسط للتناسب بالنسبة إلى نموذج تنبؤي ما، مع متغير تابع مستمر، هو نموذج R^2 أو R^2 المعدلـ نسبة التباين المفسر بواسطة النموذج. ولكن عندما يكون نموذج تنبؤي له متغير تابع ثنائي من قبيل نعم/ لا، أو صفر/ واحد، نحتاج إلى طريقة مختلفة لتقييم التناسب. وتدعى التهيئة الأكثر شيوعاً لتقييم التناسب، مصفوفة ارتباك (Confusion Matrix)، التي هي مجرد جدول ثنائي. كما أن مصفوفة الارتباك تخبرنا بمدى دقة أداء النموذج التنبؤي الذي شكلناه في تصنيف الحالات. إنه يقارن النتيجة التي يتم تنبؤها (نعم/ لا) بالنتيجة المرصودة أو الحقيقية (نعم/ لا).

وتوجد في مصفوفة ارتباك حقيقية، أعداد في الخانات الأربع؛ إذ مثّلنا – في المثال المعروض في الجدول رقم 4.3 الأعداد من قبيلn_{11} ، n_{22}، n_{21}، n_{22}، فقط من أجل الإشارة إلى خانات محددة. ولاحظ ما يلي في علاقته بهذا الجدولـ

- تظهر الترصدات التي يتم تنبؤها أو تصنيفها بشكل صحيح على الخط المائل للمصفوفة: تلك الحالات التي تم تنبؤها سلباً ورصدت سلباً في حقيقة

الأمر، زائد تلك الحالات التي تم تنبؤها إيجاباً، ثم رصدت إيجاباً. أما بالنسبة لنموذج دقيق، فإن معظم حالاته يجب أن تظهر بشكل مثالي على الخط المائل.

- أما نسبة الترصدات المصنفة بشكل صحيح بواسطة النموذج، فهي: $(n_{11} + n_{22}) / (n_{11} + n_{12} + n_{21} + n_{22})$.

- لكن في المقابل، تنقل المنشورات بشكل مألوف تصنيفاً عاماً لمعدل الخطأ، عوض $(n_{21} + n_{12}) / (n_{11} + n_{12} + n_{21} + n_{22})$.

- وتنقل بعض المقالات قياساً يدعى الحساسية (Sensitivity)، ويُعرِّف بـ $n_{22} / (n_{21} + n_{22})$.

- وتنقل بعض المقالات أيضاً قياساً يُعرف بـاسم الخصوصية (Specificity)، وتُعرِّف بـ $n_{11} / (n_{11} + n_{12})$.

- ويُعرِّف معدل إيجابي كاذب بنسبة الصور الإيجابية المتنبّأَة التي كانت في الحقيقة سلبية: $n_{12} / (n_{12} + n_{22})$.

- يعرف معدل سلبي كاذب بنسبة الصور السلبية المتنبأة التي كانت في الحقيقة إيجابية: $n_{21} / (n_{11} + n_{21})$.

وفي جميع الحالات التنبؤية، هناك مبادلة (Trade-Off) لا مفر منها في التنبؤ بين معدل إيجابي كاذب، ومعدل سلبي كاذب، أو بين الحساسية والخصوصية. وإن عملية تقليص المعدل الإيجابي الكاذب سيزيد بالضرورة من عدد المعدلات الإيجابية الكاذبة. وفي المقابل، إن تجنب المعدلات الإيجابية الكاذبة يعني أن نسبة المعدلات السلبية الكاذبة سترتفع.

الجدول رقم 4.3: مصفوفة ارتباك.

		المحصلة المتنبّأة	
		سلبي (0)	**إيجابي (1)**
المحصلة الحقيقية	سلبية (0)	n_{11}	n_{12}
	إيجابية (1)	n_{21}	n_{22}

استخدام مصفوفة ارتباك من أجل قرارات التصنيف

يمكن لنموذج انحدار لوجيستي ذي متغير تابع صفر/ واحد، أن ينقل بالنسبة إلى كُلّ ترصد أو حالة في مجموعة البيانات، الاحتمال المتنبًّأ: Y = 1. وستأخذ تلك الاحتمالات المتنبّأة مجموعة مستمرة من القيم من صفر إلى واحد. ولكن أين يتعين على باحث ما تعيين احتمال «الشريط» أو العتبة، بحيث يفترض أن يكون فوق هذه العتبة ترصداً بقيمة Y = 1، في حين يُتوقع أن يكون تحت هذه العتبة، ترصداً بقيمة Y = 0؟

ففي ببرمجيات إحصائية، عادة ما يتم تعيين الشريط في $p = 5$. ولهذا، يعالج برنامج انحدار لوجيستي كُلّ الترصدات باحتمال متنبّأ 5.أو أكبر من ذلك، باعتبارها تنبؤات بقيمة Y = 1، كما يعالج كُلّ الترصدات باحتمالات تقل عن 5.باعتبارها تنبؤات بقيمة Y = 0.

ومع ذلك لا يصح للمرء أن يفترض – بالنسبة إلى معظم قرارات العالم الحقيقي – أن تكون قيمة 5. باعتبارها سقفاً، القيمة الأفضل للتنبؤ، لأنه في الغالب هناك عدم التناسق ما بين «تكلفة» التنبؤات الإيجابية الكاذبة وبين تكاليف التنبؤات السلبية الكاذبة. وقد يكلفك تنبؤ إيجابي كاذب أكثر بكثير من تنبؤ سلبي كاذب أو العكس بالعكس، ويجب أن يُشْعِر ذلك نقطة اتّخاذ قرارك.

كيف يا ترى يتسنى لك تحديد نقطة اتّخاذ القرار، التي تعد الاحتمال المتنبأ المحصل عليه من نموذجك حيث تصنف حالة ما باعتبارها Y= 1؟ وهذا مثل واحد من أمثلة المنطق يتم استخدامه. ولنأخذ حالة بنكية حيث ضرورة اتّخاذ قرار بشأن

تقديم أو عدم تقديم قرض بملغ 5,000$ (الجدول رقم 5.3). لقد تم تشكيل النموذج للتنبؤ بما إن كان شخص ما سيفي بالتزاماته (أي لا يرجع المبلغ الذي اقترضه). لندعو P_D احتمال عدم وفاء المقترض بتعهداته، ومن ثم $1- P_D$ يشير إلى احتمال وفائه بتعهداته، فيدفع ما عليه من قروض.

الجدول رقم 5.3: إضافة اعتبارات التكلفة/ المنفعة إلى مصفوفة الارتباك.

	القرار (التنبؤ انطلاقاً من النموذج)	
	تقديم القرض	سحب القرض، الخوف من عدم الوفاء بالتعهدات
P_D الوفاء بالتعهدات	\$0	-\$5,000
$1-P_D$ عدم الوفاء بالتعهدات	- \$200	+\$200

وفي كُلّ خانة، توجد تكلفة القرار بخصوص كُلّ محصلة. ولا بُدَّ من اشتقاق هذه المعلومة من خارج نموذج التنبؤ، من شخص يدرك سياق العالم الحقيقي الذي يشتغل ضمنه النموذج؛ فإذا ما أشار نموذجك التنبؤي إلى عدم وفاء طالب قرض ما بالتزاماته، ومن ثم حجبك القرض عنه، فلن تخسر أي شيء؛ وهكذا، سيكتب 0$ في أعلى يسار الخانة داخل الجدول. وإذا تنبأ نموذجك بوفاء الشخص بتعهداته، ومنحته قرضاً على هذا الأساس، ولكنه في نهاية المطاف، أخل بالتزاماته، فستخسر قيمة 5,000$- التي اقترضتها، ومن ثم، فستكتب 5,000$- في أعلى يمين الخانة. وإذا ما تنبأ نموذجك بعدم وفاء الشخص المقترض بالتزاماته، ورفضت منحه القرض في وقت يمكن لهم مع ذلك، تسديد قرضهم، فستفوت على نفسك فرصة ربح فائدة تقدر بـ 200$ (وستكتب -200$، في أسفل يسار الخانة). وأخيراً، إذا تنبأ النموذج بعدم وفاء الشخص بالتزاماته، ومع ذلك منحته القرض، فستحقق ربح فائدة تقدر بـ 200$ (أسفل يمين الخانة).

إن القيمة المتنبأة هي: $(I- P_D) 200 - 5000 (P_D) + 200 (I- P_D) - 200 - 0(P_D)$.

ومن ثم، تكون نقطة القرار حيث كانت المحصلة التالية:

$$-200 (1-P_D) = -5000 (P_D) + 200 (1 - P_D)$$

وإذا ما أعدنا ترتيب هذه المعادلة وحلها، فسنحصل على: P_D=0.74. وتكون نقطة القرار المربحة هي منح القرض. (توقع عدم وفاء الشخص بالتزاماته) بالنسبة إلى أي قيمة متنبأة $P_D = 0.74$. أو أكبر من ذلك. لاحظ كيف يختلف هذا عن فكرة افتراض لزوم تصنيف أي احتمال يزيد عن 0.5، باعتباره إخلالاً بتعهدات، كما تنقل ذلك مصفوفة الارتباك بالنسبة إلى معظم برمجيات الانحدار اللوجيستي.

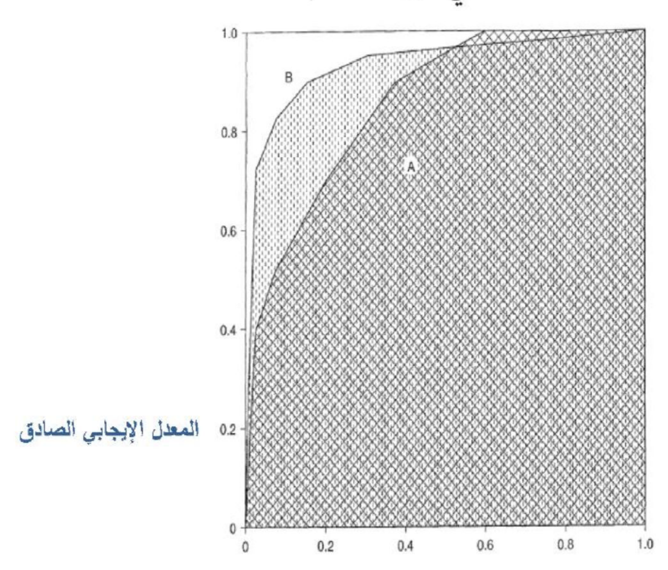
ولمزيد من الاطّلاع على ضمّ اعتبارات التكلفة في نماذج التصنيف، انظر عمل (Witten, Eibe, and Hall (2001, 163). وعادة ما تكون إضافة اعتبارات التكلفة إلى مصفوفة الارتباك من أجل البتّ في النقطة الفاصلة، مباشرة عندما تكون للتكاليف والنفقات قيمة نقدية مباشرة. ولسوء الحظ، فإنه يصعب التعبير أحياناً عن المبادلة بين قيم إيجابية كاذبة وقيم سلبية كاذبة أو بين الحساسية والخصوصية، بهذا الشكل. وإن البتّ في مكان وضع الحدّ الفاصل بالنسبة إلى اختبار صحة تشخيص جديد محفوف بالصعوبة، بما أن المرء مطالب بتحقيق التوازنات بين الاضطرابات التي تحدث عندما يقال للمريض خطأً إنه يعاني من بعض المشاكل الطبية الخطيرة، وبين نتائج الإخفاق في تحديد تلك المشاكل عندما يكون قائماً في حقيقة الأمر.

منحنيات خاصية التشغيل المتلقي باعتبارها مقاييس مطابقة

إن منحنى خاصية التشغيل المتلقي (Receiver Operating Characteristic (ROC))، طريقة مرئية للبتّ في أفضل النماذج المستخدمة في تصنيف الحالات. ويستخدم في سياقات تكون فيها المحصلة صفر/ واحد أو نعم/ لا، كما يقدم فيها نموذج ما، احتمالاً متنبأً من احتمالات «نعم» أو Y = 1 بالنسبة إلى كُلّ حالة. (وعلى مستوى التنقيب في البيانات، تعد هذه مصنفاً ثنائياً (Binary Classifier). إن العديد من الاختبارات الطبية هي مصنفات ثنائية، مثلاً). ويخُطُّ منحنى خاصية التشغيل المتلقي، معدل الإيجابيات الصادقة (خصوصية) على المحور Y، مقابل الإيجابيات الكاذبة (1 - خصوصية) على محور X. إنها تصف - إذن - التبادلية (تجارية) بين الإجابيات الصادقة (الربح)، وبين الإيجابيات الكاذبة (التكلفة) - انظر الشكل رقم 3.3.

وفي هذا الشكل، يعد النموذج الممثل بخط B، عموماً متفوقاً في التصنيف على ذلك الممثل بخط A. ولكن يمكن أيضاً أن نرى أن لنموذج A أداء تنبؤياً أفضل (Fawcett 2006)، حيثما كان المعدل الإيجابي الكاذب عال جداً (أكبر من 0.6).

وإن منحنى خاصية التشغيل المتلقي، غالباً ما يُستخدم لفهم دقة الاختبارات التشخيصية لمرض ما، مثل فحص الدم. وبالإضافة إلى الاحتمال المتنبأ بشأن إصابتهم بالمرض، المحصّل عليه انطلاقاً من فحص الدم، يحتاج المرء إلى معلومة موضوعية منفصلة تثبت ما إذا كان الشخص فعلاً مصاباً بالمرض. ويدعى هذا الأخير «معايير الذهب» (Gold Standard) في أدبيات الطبّ.



الشكل رقم 3.3: أمثلة من منحنيات خاصية التشغيل المتلقي(Fawcett 2006).

ويلي نموذج مثالي بشكل وثيق محور Y على الجانب الأيسر وبعدها يتحول بالموازاة إلى المحور X. ويقترب – قدر الإمكان – من أعلى يسار ركن مخطط منحنى خاصية التشغيل المتلقي. إنَّ المنطقة تحت هذا المنحنى هي حوالي 1. ويمكن أن يلي نموذج سيء ما، الخط ذا 45 درجة: وليس هذا أفضل من حظ، أما المنطقة التي هي تحت هذا الخط، فتبلغ 0.5. وهذا الاختبار التشخيصي، لا يزودك بأي شيء مفيد.

ولتلخيص هذا القسم، غالباً ما يشكل التنقيب في البيانات نماذج تنبؤية، ويريد المختص في التنقيب في البيانات طريقة من الطرق من أجل تقييم دقة نموذج معين.

أولًا: يطبق المختص في التنقيب في البيانات نموذجاً تنبؤياً، الذي اشْتُقَّ من بيانات التدريب، على بيانات أخرى (حالات أو ترصدات أخرى) خصصت باعتبارها بيانات اختبار. ويستخدم عالم التنقيب في البيانات مصفوفة ارتباك أو منحنى خاصية التشغيل المتلقي لفهم دقة النموذج في تنبؤ هدف ما. ويتمثل واحد من القياسات المهمة للتطابق أو دقة التصنيف في نسبة الترصدات المصنفة بشكل صحيح (أو في المقابل، معدل الخطأ الشامل). ولكن الباحث غالباً ما يرغب في تحديد المعدلات الإيجابية الكاذبة، والمعدلات السلبية الكاذبة، وفي بعض الأحيان يستخدم هذه المعلومة، إلى جانب بيانات التكلفة، للبتّ في خفض القيمة الأكثر ملاءمة واستخدامها مع الاحتمال المتنبأ لدى تصنيف الحالات.

تحديد تفاعلات إحصائية وتأثير عدم التجانس في التنقيب في البيانات

وتتجلى إحدى الرغبات الأساسية في نموذج الانحدار التقليدي، في تطبيق نمط الترابط أو الارتباط (Correlation) نفسه على كُلّ الترصدات في مجموعة بيانات ما. وعندما تحدث الحالة العكسية - أي عندما تضمّ مجموعة بيانات، مجموعات ترصدات ذات علاقة مختلفة جداً بين المتغيرات - يكون بإمكان نماذج الانحدار إنتاج معامِلات مضللة للغاية. ويعرف هذا - بشكل عامِّي - بمشكل «التفاح والبرتقال»، أو على نحو أكثر تقنية بالتأثيرات غير المتجانسة (Heterogeneous Effects). على سبيل المثال، إذا كانت العوامل التي تتنبأ بنسبة التخرج بين طلبة كليات المجتمع، مختلفة للغاية عن العوامل المترابطة بالتخرج بالكليات الانتقائية، ذات التكوين الممتد لأربع سنوات، فسيسفر تقدير نموذج إحصائي وحيد بالنسبة إلى مجموعة بيانات تضم النوعين من الطلبة معاً، عن نتائج مضللة.

إن المشكلة لا تكمن في وجود مجموعات مختلفة داخل مجموعة بيانات؛ لكون استمرار هذا الأمر على هذا النحو بشكل دائم، وإنما المشكلة تنشأ عندما يكون لبعض المجموعات الفرعية أو لبعض (Clusters) حالات داخل مجموعة بيانات،

أنماط مختلفة جداً من الترابط بين متغيرات ما، أكثر من مجموعات أخرى. وتعرف إحدى الأمثلة المثيرة بـ مفارقة سيمبسون (Simpson's Paradox)، (وأحياناً بمفارقة يول - سيمبسون (Yule-Simpson's Paradox))، أو مفارقة الإدماج، أو المفارقة العكسية (Blyth 1972). وقد تظهر مجموعتان من الترصدات في مجموعة بيانات، علاقة إيجابية بين متغيرين، XوY. ولكن عند تحليل مجموعتين معاً في النموذج نفسه، فإن اتّجاه العلاقة بين X و Y تعكس الاتّجاه. وقد يبدو ارتباط X سلبياً بـ Y.

ويوجد مثال في الجدول رقم 6.3، يقدم محصلات تجربة طبية مفترضة، تم القيام بها في موضعين (موقعي A و B)، بحيث تشمل منح بعض المرضى معالجة جديدة تجريبية، ومنح آخرين المعالجة المعيارية. وتقارن الخانتان الأوليتان، محصلات المعالجة التجريبية، والمعالجة المعيارية للتجربة بشكل عام عبر الموقعين؛ فظهر بجلاء أن معدل البقاء على قيد الحياة كان أقل بكثير بين أولئك الذين يتلقون معالجة تجريبية. وإذا ما فحصنا هذين الخانتين بمفردهما، فسنخلص إلى أن المعالجة التجريبية أسوء بكثير من المعالجة المعيارية، ومن ثم، ضرورة التخلي عنها بالمرة.

الجدول رقم 6.3: صياغة بليث لمفارقة سيمبسون.

	إجمالاً		الموقع A		الموقع B	
	معالجة معيارية	معالجة تجريبية	معالجة معيارية	معالجة تجريبية	معالجة معيارية	معالجة تجريبية
الرقم الإجمالي	11,000	10,100	1,000	10,000	10,000	100
عدد الموتى	5.950	9,005	950	9,000	5,000	5
عدد الأحياء	5.050	1,095	50	1,000	5,000	95
معدل البقاء على قيد الحياة	46٪	11٪	5٪	10٪	50٪	95٪

المصدر: Blyth, 1972.

ولكن، عندما ننتقل إلى فحص الخانات الأربع من ناحية اليمين، نجد في كُلّ موقع من هذين الموقعين الفرديين، أن معدل البقاء على قيد الحياة، كان أكثر بكثير بين أولئك الذين يتلقون العلاج التجريبي. وتقترح هذه الملاحظة أن العلاج التجريبي أكثر فاعلية من العلاج المعياري. فكيف - إذن - السبيل إلى التوفيق بين هذا وبين

البيانات المجمّعة (Aggregated)؟ والجواب عن ذلك يكمن في تدبير التقنية التجريبية - في أغلب الأحيان - في موقع ذي معدلات منخفضة من معدلات البقاء على قيد الحياة بالنسبة إلى المجموعتين معاً؛ في حين كانت تدار التقنية المعيارية - على نحو غير متكافئ - في موقع ذي معدلات أكبر بكثير من معدلات البقاء على قيد الحياة. وعندما يتم مزج البيانات، تختفي النسبة العالية من معدلات البقاء على قيد الحياة المحصل عليها في المجموعات التجريبية؛ أو بتعبير آخر، إن العلاقة السلبية المرصودة ذات المتغيرين بين تلقي العلاج التجريبي، واحتمالات البقاء على قيد الحياة، تتجه اتّجاهاً عكسياً، عندما نكيّف الموقع الذي يتلقى فيه الشخص العلاج.

وتوجد حالة أقل حدة، ولكنها أكثر شيوعاً، تحدث عندما يبدو معامل انحدار مرصود بالنسبة إلى متنبئ من متنبئات X، صغيراً أو عديم الدلالة إحصائياً. ويحدث هذا أحياناً، بسبب ارتباط XبـY ارتباطاً قوياً، بالنسبة إلى مجموعة واحدة أو تجميع من الحالات داخل العينة، في حين قد تنعدم العلاقة، أو أي علاقة سلبية مع Y، بالنسبة إلى مجموعة أخرى ذات المتنبئ X نفسه. وإن استخراج متوسط هذين التأثيرين - كما يفعل الانحدار عند تحليل العينة بأكملها - يفضي إلى معامل صغير على نحو مضلل.

وغالباً ما تكون مجموعات البيانات غير متجانسة على هذا النحو، غير أن الباحث لا يدرك عادة، المجموعات الفرعية أو تجميعات الحالات مسبقاً، ومن ثم، تظل مشكلة «التفاح والبرتقال» مشكلة متوطنة. ونتيجة لذلك، تقتضي خطوة أولية في تحليل التنقيب في البيانات، الرغبة في تحديد المجموعات أو تجميعات الحالات، بغية تمكّن باحث ما - بعد ذلك - من إدارة إما تحليلات منفصلة بالنسبة إلى كُلّ مجموعة متميزة على حدة، أو إضافة شروط تفاعل تُنَمذِجُ انحدارات مختلفة بالنسبة إلى كُلّ مجموعة أو تجميع (Melamed, Breiger, and Schoon 2013).

ويمكن لتقنيات تجميع عديدة من تقنيات التنقيب في البيانات، تحديد تجميعات الترصدات ذات العلاقات غير المتجانسة بين متغير أو مزيد من متغيرات X، وبين محصلة من محصلات Y. وقد طور روبرت هاراليك (Robert Haralick)، وزملاؤه طريقة يستعمل فيها تجميعاً متشعباً خطياً ولا خطياً (Haralick and Harpaz 2007).

كما يقدم ميلاميد وبريغر، وشون، حلاً آخر، مستخدماً تجزؤ (Decomposition) القيمة المنفردة. ولسوء الحظ، إن هذه التقنيات ليست متاحة لحدِّ الساعة، في أي رزمة من رزمات برمجيات التنقيب في البيانات الأساسية.

ويستخدم حلّ ثالث أكثر سهولة، تقنية تعرف بانحدار الطبقة الكامنة (Latent Class Regression)، أو نماذج تجميع الطبقة الكامنة (Latent-Class Cluster Models). وتم استخدام تعبير طبقة كامنة لعدم إمكانية تحديد المجموعات غير المتجانسة داخل مجموعة بيانات، بواسطة متغير مقيس واحد (Single Measured Variable). (وإذا اختلف الرجال عن النساء في نموذج انحدار ما، أو إذا أظهر المبحوثون أو المستطلعون (Respondents) الشباب نمط ترابط مختلف بين المتغيرات، فبإمكان تحديد ذلك بشكل سهل نسبياً، ما دام أن هذه المتغيرات، هي متغيرات مرصودة واحدة). وعندما يتم تحديد المجموعات الفرعية في البيانات بطرق أكثر تعقيداً، نتصور أن للمجموعات الفرعية قيماً مختلفة على متغير غير مرصود (ومن ثم، فهو «كامن»). فكيف يحدد المرء - إذن - هذه المجموعات الفرعية؟

تقدم الابتكارات الإحصائية (Statistical Innovations)، رزمة برمجيات، تُدعى «الذهب» الكامن (Latent GOLD)، التي تنجز هذا النوع من التحليل على نحو واضح وسهل. وتوصف البرمجيات على موقعها الإلكتروني: (http://statisticalinnovations.com/products/latentgold.html/).

وغالباً ما يدعى هذا الموضوع بين علماء الإحصاء بنمذجة المزيج المنتهية (Finite Mixture Modeling)، كما حدث في العقد الأخير، وهناك تقدم كبير في تطوير هذه التقنية. وكتب (Collins and Lanza 2009)، كتاباً مفيداً حول الأفكار الإحصائية التي تحملها هذه الطريقة.

وبعد تحديد التجميعات المتميزة أو مجموعات الترصدات داخل مجموعة بيانات ما، قد يقرر باحث من الباحثين تحليل كُلّ التجميعات في نماذج منفصلة. ويمكن للتجميعات - بدلاً من ذلك - أن تمثل بواسطة متغير اعتباري (Nominal Variable)، ومصطلحات جديدة تنضاف إلى النموذج الذي يمثل تفاعلات بين

متنبئات تجميع، وبين متنبئات خاصة. كما يمكن لنموذج واحد، المؤلف من هذه المصطلحات التفاعلية أن يُحسبَ للأخذ بعين الاعتبار تأثيرات المجموعات غير المتجانسة.

غابات تعبوية وعشوائية

عادة ما تكون النتيجة النهائية في النموذج الأصلي الإحصائي التقليدي، انحداراً وحيداً أو نموذجاً مماثلاً، يلخص العلاقات في مجموعة بيانات ما. قد يمرّ ذلك النموذج عبر سلسلة من التحسينات والتعديلات، ولكن في النهاية، يمثل نموذج بمفرده أفضل ما يمكن لباحث من الباحثين الإتيان به.

وفي المقابل، يتْبَع - في الغالب - تحليل من تحليلات التنقيب في البيانات، منطقاً مختلفاً، مولّدين العديد من النماذج التنبؤية المختلفة، ومزج نتائجها لتقديم أفضل تنبؤ ممكن، وهي عملية تعرف في مجال التنقيب في البيانات باسم تعليم طاقم منسجم الأجزاء (Ensemble Learning) (Berk 2006). وثمة استراتيجيات بديلة داخل التنقيب في البيانات لخلق هذه النماذج المتعددة ومزجها، ومنها استراتيجية التعبئة، (وينبغي عدم خلطها بتوزيع الخانات (Binning))، التي تتعامل مع مجموعة بيانات كما لو كانت ساكنة، وليست عينة. إنها تستمد عينات عشوائية متعددة مع استبدال (With Replacement)، من مجموعة البيانات. ويناسب تطبيق التنقيب في البيانات، نموذجاً لكُلّ عينة عشوائية من تلك العينات، وانطلاقاً من ذلك النموذج، يَحسبُ قيمة متنبّأة لمتغير النتيجة (Outcome Variable)، بالنسبة إلى كُلّ حالة أو ترصد. ويمكن إيجاد تنبؤات مأخوذة من تلك النماذج المختلفة، بغية تحقيق أفضل تنبؤ ممكن، إما لمجموعة البيانات المنفصلة، أو للترصدات الجديدة المستخلصة من العينة.

وثمة مقاربة ذات الصلة، تعرف باسم الغابات العشوائية (Random Forests)، تستعمل لجمع نتائج أشجار قرار متعددة. وتتمثل الفكرة الرئيسة في توليد نماذج شجرة متعددة، وإيجاد معدل نتائجها للحصول على أفضل تنبؤ، كما يتجلى المظهر الجديد للغابات العشوائية في فرض الباحث مجموعة فرعية مختلفة من التنبؤات،

لضمها في كُلّ نموذج، ومن ثم عدم تمكين أي نموذج بنية متطابقة أو مضمون مطابق للنموذج السابق. وبعد ذلك، يتم مزج التنبؤات المتنوعة المحصل عليها من تلك النماذج المتعددة للحصول على أفضل تقدير.

إن الغابات التعبوية والعشوائية هي إجراءات اختيارية داخل (JMP) التي تنطق «غامب برو»، والعديد من مجموعات (Suites) التنقيب في البيانات التي نوقشت آنفاً. سيتم تقديم أمثلة عنها في فصل لاحق.

إن إحدى الأسباب الجوهرية لممارسة التنقيب في البيانات وفي تقييم نماذج متعددة، وإيجاد متوسطات نتائجها، تكمن في إمكانية أن يكون في بعض الحالات، بناء النموذج تابعاً للمسار (Path Dependent). وفي أنواع متعددة من نماذج التنقيب في البيانات، تفحصُ خوارزمية ما كُلّ سمة، لاستكشاف النموذج الذي يشكل التنبؤ الوحيد الأكثر قوة لهدف ما؛ فتحتفظ بالنموذج الأكثر قوة، وتعيد البحث في التنبؤات المتبقية، لانتقاء المتنبئ الثاني من حيث القوة، وهكذا بالنسبة إلى العديد من التكرارات (Iterations) إلى أن تنتقي مجموعة من السمات أو المتغيرات التي تعظم بشكل جماعي، القوة الشاملة للنموذج.

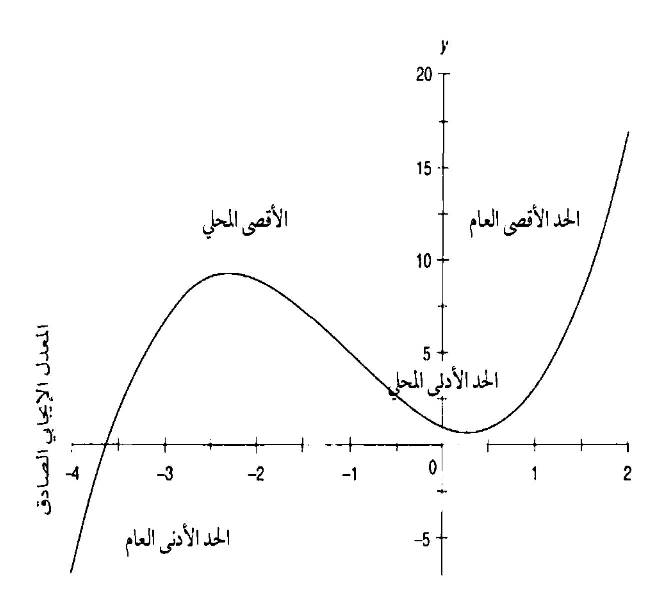
وهذه طريقة مستعملة على نطاق واسع، ومناسبة في اختيار المتغير أو السمة، على الرغم من أنَّ لها شركاً محتملًا. لقد اختارت هذه الخوارزمية - ذات مرة - متنبئاً أولاً، لإدخاله في النموذج؛ مما زاد من احتمال اختيار بعض المتغيرات باعتباره المتنبئ الثاني، مقارنة بمتنبئات أخرى، فمن غير المرجح، مثلًا أن تختار الخوارزمية متغيراً ما - باعتباره المتنبئ الثاني - المرتبط بشكل كبير بالمتغير الأول الذي اختارته، بما أن إضافة متنبئ ثانٍ وثيق الصلة، لن يحسّن القوة التنبؤية كثيراً. وبتعبير آخر، إن اختيار المتغير الأول - إلى حدّ ما - يحدد مساراً من المسارات بالنسبة إلى التكرارات المتبقية للبرنامج، ومن تبعية المسار.

وتقتضي تبعية المسار تجاهل بعض المتنبئات القيمة أو إزالتها في أي نموذج من النماذج. وبهذا، سيكون من المنطقي تقييم العديد من النماذج المقيدة باختيار متنبئات مختلفة، ومن ثم تجنب إمكانية تجاهل بعض المتنبئات، وهذا ما تنجزه الغابات العشوائية.

ويضم منطق ذو صلة، إجراءات التنقيب في البيانات التي تشمل تقديرات تقريبية، تلتئم في حلّ أمثل. واستناداً إلى موقع «التخمين» الأول، إن إحدى نقاط ضعف هذه الخوارزمية، يتجلى في إمكانية التئام هذا البرنامج أحياناً، في حلّ محلى أمثل، الذي لا يعد في حقيقة الأمر، الحلّ الأفضل إجمالياً.

ويفهم هذا بيانياً (Graphically)؛ إذ إن في الشكل رقم 4.3 يمثل محور Y قياس خطأ ما، ومن ثم، فالبرنامج يبحث عن حلّ يتميز بأدنى قيمة ممكنة على محور Y. كما يمثل محور X، قيمة مَعْلم ما تم تقديره. أما المنحنى، فيمثل المسار الذي يمكن لبرنامج ما اتباعه في البحث عن حلّ من الحلول، المتمثل في أفضل تقدير لـ X. وإذا كان التخمين أو التقدير الأول للبرنامج موجوداً على الجانب الأيسر من الرسم البياني أو المخطط (Diagram) – عند قيمة منخفضة لـ X – فإن عملية مكررة، تختار كُلّ حلّ متعاقب منخفض انخفاضاً طفيفاً على محور Y، سيتبع الخط الأسفل إلى أن تصل إلى الحدّ الأدنى العام (Global Minimum)، أي أفضل جواب ممكن. ولن تحرك الخط إلى الأعلى عندما يتحول الخط صعوداً، لأنها مبرمجة على مواصلة البحث عن قيم منخفضة لـ Y والتوقف عند عجزها عن إيجاد قيمة أقل انخفاضاً. وتتوقف الخوارزمية عند القيم الأدنى لـ Y، أي الحدّ الأدنى العام، حيث تقدير X هو حوالي – 4.

ولكن إذا كان التخمين أو التقدير الأول للبرنامج موجوداً على الجانب الأيمن من المخطط – عند قيمة مرتفعة لـ X (1.5 مثلاً) – فستتجه العملية المتكررة اتّجاهاً منحدراً (Down-Slope) نحو حدّ أدنى محلي (Local Minimum). ولأن الخوارزمية دائماً ما تحاول تخفيض Y، فلن ترجع إلى الخلف في الجانب الأعلى للخط، بعد بلوغها أول نقطة منخفضة حوالي 0.3 =X، ومن ثم، ستفقد «الوادي» (Valley) المقبل حيث إقامة الحدّ الأدنى العام. وتستقر بعدها في الحدّ الأدنى المحلي (حوالي X=0.3)، «معتقدة» خطأ في كون ذلك هو أفضل حلّ، أي تقدير X الذي يقلص Y إلى الحدّ الأدنى.



الشكل رقم 4.3: الحد العام والحد الأدنى المحلي.

وتتجلى إحدى السميات المضادة لهذا الشكل ذي الطرق المتكررة، في تقدير نماذج مختلفة متعددة - بحيث يبدأ كُلّ واحد منها عند نقطة بداية مختلفة جداً (تخمين أو تقدير أولي) - وفي تجميع التنبؤات من كُلّ هذه النماذج المختلفة لتحديد تنبؤ نهائي لشخص ما. ولن يمنع هذا الإجراء بعض الحلول من أن تكون مثالية فرعية (لأنها استقرت في حدّ أدنى محلي)، ولكنه يضمن وجود حظوظ أخرى عديدة لبلوغ الحلّ الحقيقي أو المثالي (الحد الأدنى العام)، وستكون هي المهيمنة.

إن «غامب برو» (JMP Pro)، ورزمات برمجية أخرى من التنقيب في البيانات، تسأل المستخدِم عن عدد نقاط البداية المستخدمة. وبعد ذلك، تدير نماذج منفصلة، تبدأ عند نقاط بداية مختلفة للغاية، لضمان عدم الانخداع بحدود أدنى محلية. وتتمثل التكلفة - عادة - في ضرورة إدارة نماذج عديدة عوض نموذج واحد، مما قد يستغرق وقتاً كثيراً لمعالجة مجموعات بيانات ضخمة.

ولتلخيص هذا القسم حول الغابات التعبوية والعشوائية، اكتشف الباحثون الذين

يشتغلون داخل النموذج الأصلي للتنقيب في البيانات، تحليل البيانات عدة مرات، مستخدمين عينات مختلفة قليلاً، أو مجموعات مختلفة من المتنبئات، أو نقاط بداية مختلفة. وكل تحليل فردي يقدم تقديراً ما، ويوجد التنبؤ الأكثر قوة ودقة، في مزج تلك التقديرات. ولاستخدام قياس ما، نقول إن قرار اللجنة هذا، ويُزعم أن «التصويت»، أو إيجاد المعدل، أو في بعض الأحيان، مزج نماذج متعددة (معروفة أيضاً بتعليم طاقم منسجم الأجزاء)، يقدم تنبؤاً أكثر دقة من الاعتماد على نموذج أو تحليل واحد. ومع ذلك، يقوم تكرار تحليلات على هذا النحو، ومزج - بعد ذلك - نتائجه، على حواسيب فائقة السرعة، وواسعة، إذ لها القدرة على حساب نماذج عدة مرات، وهو شرط ضروري للعديد من طرق التنقيب في البيانات.

محدودية التنبؤ

نشر ممارسان بارزان في التنقيب في البيانات، أفضل الكتب مبيعاً حول محدودية هذا التنقيب في البيانات والتنبؤ. نسيم نيكولاس طالب (Nassim Nicholas Taleb)، وهو محلل مالي، ومنمذِج إحصائي، يعد مؤلف كتاب *The Black Swan* (2005)، وكتاب *Fooled by Randomness* (2007)؛ كما ألف نات سيلفر (Nate Silver)، وهو مطور برمجية تنبؤية في «البيسبول»، ومحلل رائد في استطلاعات الانتخابات (انظر مدونة *Five Thirty Eight* لـ *New York Times*)، كتابـ *The Signal and The Noise* (2012).

ويقدم المؤلفان كلاهما نقاطاً تحذيرية، مثلـ

- ليس لكُلّ الظواهر الطبيعية أو الاجتماعية بنية أساسية، يمكن استكشافها. عموماً، كلما كانت نسبة الصوت في الإشارة (Signal)، ازدادت نسبة تضليل الإفراط في التدريب لمختصي التنقيب في البيانات. وقد «ينخدعون بالعشوائية»؛ فيرون سراباً، أو يجدون بنية غير موجودة. (من أجل ذلك، فإن الصلاحية المتبادلة، والمضاعفة، مهمتان جداً).
- إن النُّظم الدينامية المترابطة جداً، تتأثر بأسباب متعددة، بحيث يمكن إثارة بعضها، حلقات تغذية راجعة، قادرة على إنتاج تحول غير متوقع على نطاق واسع. وكان العرض بعنوان

- «Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas»، عرضاً شهير، يتناول نظرية الفوضى (Chaos Theory)، وهي نظرية تُعنى بالنظم اللا خطية التي تُنسَبُ إلى إدوارد لورينز (Edward Lorenz). وكما يقترح المجاز، يمكن لتحول ما على نطاق صغير في مكان ما، إثارة نتائج على نطاق واسع جداً في مكان آخر. وقد يفهم من هذا المجاز، استحالة نَمْذَجَةِ النظم اللا خطية بنجاح على الإطلاق. وفي المقابل، يمكن تنبؤ النظم اللا خطية مثل الطقس في حدود معينة (حسبما يرى سيلفر (Silver)) ، ولكن فقط ضمن إطار زمني قُبَيْل الحدث المتنبأ. وإن التنبؤات التي تمت في وقت سابق من هذا، لن تكون دقيقة بشكل كامل. وبتعبير آخر، لا يستطيع المرء اقتفاء أثر أي إعصار حقيقي في رفرفة أجنحة الفراشة (أثر الفراشة). وبإمكان المرء تصور العلاقة، ولكنه يعجز عن نمذجة ذلك بالأمس البعيد. ومع ذلك، يمكن تنبؤ الإعصار بشكل دقيق، انطلاقاً من أدلة يتم جمعها قبل بضعة أيام من حدوثها.

- وتُظهر أنظمة طبيعية أخرى انتظامات (Regularities)، ولكنها بعيدة عن فهمنا التنبؤي الحالي. وإن توقيت الزلازل وشدتها، مثل من الأمثلة التي يصفها سيلفر، إذ يمكننا تحديد بعض الأنماط حول حجم الزلزال، ولكن لا يمكننا التنبؤ – على نحو دقيق – بزمن حدوث الزلازل الكبيرة. إن النمط – إذا وجد واحداً أصلاً – بعيد المنال.

- وتعد بعض الظواهر الاجتماعية غير مناسبة للتنبؤ، لأن الفاعلين المشاركين فيها، يفحصون بيئتهم، ويردون على أية تلميحات من تلميحات التغيير، انطلاقاً من الوضع الراهن. وفي ظل تلك الظروف – وبمجرد إدراك تلميح من حركة – مثلاً، أخذت أسعار الأسهم في الارتفاع أو الانخفاض – سيتوقع العديد من الناس أن السوق يُحول اتّجاهه، فينضمون إلى الفريق الرابح. وقد يصير – إذن – للحركة توجه نحو إشباع الذات (Self-Fulfilling Prophecy)، بما أن مزيداً من الناس تشتري الأسهم أو تبيعها. وفي هذه السياقات المتقلبة، يعد سلوك البارحة متنبئ ضعيف لتصرّف الغد، بما أن

المشاركين قد يكسبون مالياً من توقع تغيير في الاتّجاه. وفي هذا السياق، يمكن لعقلية القطيع تقويض التنبؤ.

- وفي سياقات أخرى - مع ذلك - يمكن لآراء جماعة من المواطنين منافسة تنبؤات الخبراء؛ فتنبؤات الأشخاص حول العديد من القضايا، كثيراً ما يفوق تقديرات معظم الأفراد. من أجل ذلك، غالباً ما تنجز «أسواق» المعلومات، حيث يراهن العديد من الأفراد على النتائج، ما يتنبأ به صناع القرار وتستشيرهم.
- وغالباً ما تكون لدى الخبراء ثقة مفرطة بالنفس بشأن تنبؤاتهم. وعلى نحو مماثل، يميل الأفراد الذين لهم رهان في وضع راهن ما، إلى إهمال حقيقة خطر التغيير أو التغاضي عنه.

إن أهم نقطتين حاسمتين لهذين المؤلفين - في رأيي - هما كالتالي:

- يجب على أي تنبؤ أن يكون دوماً مصحوباً باحتمال أو فاصل الثقة التي تمثل الشكّ في التنبؤ.
- وإنّ النتائج المستبعدة إحصائياً - وإن كان ذلك نادراً - تحدث بكل تأكيد، ولا بُدَّ من تنبؤها إذن، والتخطيط لها. وستحدث حالة واحدة في المليون عند نقطة ما. ويشير طالب (Taleb) إلى هذه الأحداث بعبارة البجعات السوداء (Black Swan)، ويتناول تأثيرها المدمر في الناس الذين يتخذون قرارات، مدعين عدم حدوث النتائج النادرة إحصائياً. وتكمن إحدى الأسباب وراء انهيارات السوق، وحالات الإفلاس، في ميل صنّاع القرار المَهَرَةِ كمياً، إلى التصرف كما لو أن الأحداث الأكثر احتمالاً هي فقط التي ستحدث مستقبلاً. وستنهار مشاريعهم (وثرواتهم) عندما تحدث بجعة سوداء، وهو حدث مستبعد.

البيانات الضخمة ليست بالمرة ضخمة بما فيه الكفاية

إن الصورة التي رسمناها عن التنقيب في البيانات، تقترح مزج قوة الحوسبة القاسية (Brute Computing Power) ومجموعات بيانات ضخمة جداً، تمكّن

متخصصو التنقيب في البيانات من استكشاف البنيات في البيانات التي لم يتم الإفصاح عنها، من خلال تطبيق المنهجيات الإحصائية التقليدية على مجموعات البيانات المؤلفة من عدد أصغر من الحالات. إننا ندعم هذا الطرح، غير أنه مهم أيضاً الاعتراف بمفارقة تواجه متخصصي التنقيب في البيانات باستمرار، وتشكل مشروعهم بأكمله، أي إنه حتى مجموعات بيانات العلوم الإنسانية الكبرى - مثلاً خمسة ملايين شخصاً، ملفات تعداد لعدة سنوات متاحة من مسح المجتمع الأميركي (American Community Survey)، ليست كبيرة بما فيه الكفاية لتمكننا من بحث شامل وحصري في البنية، بل إنَّ الحواسيب الكبرى والفائقة السرعة نفسها، تجد بعض المهام التجريبية صعبة المراس.

إن إحدى نتائج هذه المقارنة تتجلى في أنَّ التنقيب في البيانات، كثيراً ما يحتاج إلى وضع افتراضات مبسطة كي تكون الحلول للمشاكل ممكنة، أو لانتقاء مجموعات فرعية من المتغيرات، لأن التنقيب في البيانات نفسه، لا يمكنه التعامل مع كُلّ القياسات المتاحة في نموذج واحد. وإذا ما أخذنا بعين الاعتبار المعالجة الحسابية (Computing) الضخمة، ومصادر البيانات المتاحة، فهل يحتاج - مع ذلك - التنقيب في البيانات إلى تسوية أو اختصار النفقات أو إيجاد طرق مبتكرة للتقدير بدلاً عن قياس الأشياء على نحو مباشر وشامل؟

ويمكن لتجربة فكرة ما أن تبين مكمن الخطورة. تصور حالة من الحالات التي يكون لدينا فيها، هدف ثنائي (نعم/ لا)، أو متغير تابع، وحددنا عبر عملية من عمليات البحث الاستكشافي أن أخذ 10 متغيرات أو سمات معاً، يمنح تنبؤاً جيداً لهذا الهدف ذي ثنائية نعم/ لا. والآن، لنفرض جدلاً أن كلاً من هذه المتنبئات العشر (10) تأخذ قيماً من 0 إلى 9. (ولنأخذ، مثلاً، متنبئات مستمرة مثل العمر أو الدخل، ونقسم كُلّ واحد منها إلى 10 خانات، محولين كُلّ منها إلى متغير عادي ذي 10 قيم ممكنة).

وافتراضاً، يمكن للمرء تشكيل جدول مكون من صف (Row) لكُلّ مزج ممكن لمتغيرات أو سمات التنبؤ ذي القيم العشرة. وكل حالة من هذه الحالات أو الترصدات في مجموعة بيانات تدريب، يمكن حلها، بحيث تنتهي في نهاية المطاف في العصر

الواحد الذي يمثل قيم تلك الحالات على 10 متنبئات. وبعدما يتم حلّ كُلّ بيانات التدريب على هذا النحو، فإنه يصير بالإمكان عدّ نسبة الأجوبة بنعم بالنسبة إلى ذاك السطر.

يمكن استخدام هذا الجدول من بيانات التدريب، بعد ذلك، باعتباره نموذجاً تنبؤياً. وللتنبؤ بالهدف (نعم أو لا) بالنسبة إلى كُلّ حالة جديدة في عينة اختبار، ما على المرء إلا البحث في الجدول عن العمود الخاص الذي كان يوافق نمطاً فردياً من المتغيرات المستقلة بالنسبة إلى تلك الحالة الجديدة أو الترصد الجديد، (مثلاً، ابحث عن السطر الخاص بالرجال، الذين تتراوح أعمارهم ما بين 70,65، وأدخلهم ما بين $80k,$75k القاطنين في بريطانيا الجديدة إلى آخره، بالنسبة إلى 10 من متغيرات شخص ما. إن نسبة حالات «نعم» في ذلك العمود، ستقدم إذن الاحتمال المتنبأ «لنعم» لحالة جديدة خاصة في بيانات الاختبار، وهذه العملية من البحث يمكن إعادتها بالنسبة إلى كُلّ حالة في ملف البيانات الجديدة.

لماذا لا يكون هذا النوع من استراتيجية تنبؤية تجريبية شاملة، عملياً مع البيانات الضخمة؟ لنتأمل فيما يسميه مختصو التنقيب في البيانات حيز المقياس (Measurement Space): حجم الجدول الضروري لتمثيل كُلّ التركيبات (Combinations) للمتغيرات العشرة (10)، بحيث يحتوي كُلّ منها على 10 قيم، وقد تكون 10^{10} من حيث الحجم: عشرة مليارات عمود إجمالياً. ولنتأمل أيضاً قدر بيانات التدريب الضرورية، بحيث تكون (مثلاً) مائة حالة أو ترصد، متاحاً بالنسبة إلى كُلّ عمود داخل الجدول الذي انطلاقنا منه، يمكن عد نسبة الأجوبة بنعم بالنسبة إلى بيانات جديدة. وقد يحتاج المرء إلى مجموعة بيانات تدريب بمقدار مائة مرة لعشرة مليارات حالة - تريليون حالة - لملء الجدول بما فيه الكفاية للسماح باستراتيجية بحث تجريبية بشكل بحت. وبما أن لدينا نقص في علم الفلك، فإنه من غير المرجح إيجاد مجموعات بيانات ذات تريليون حالة. ولهذا، يعجز التنقيب في البيانات عن التعامل مع استراتيجية قياس مباشرة وشاملة بالنسبة إلى مشكلة متخيلة لعشرة متغيرات، بحيث يملك كُلّ واحد منها 10 قيم.

إن ما تنوي هذه التجربة الفكرية الإفصاح عنه، هو أن التنقيب في البيانات يواجه حالات قصور مهمة، ناتجة عن حجم بيانات التدريب والتحميل الحاسوبي. ومن جهة أخرى، إن للتنقيب في البيانات استراتيجيات عدة، تتجنب بنجاح هذه المشاكل، وتمكن طرق التنقيب في البيانات، تحليل البيانات ذات مئات المتغيرات بنجاح.

- **أولاً:** يضع التنقيب في البيانات أهمية كبيرة على عملية تقليص عدد المتغيرات التي تدخل في إطار أي نموذج، وتشمل إحدى المقاربات، انتقاء السمة (Feature Selection): عملية الفحص عبر أعداد هائلة من المتغيرات لاستكشاف المجموعة الفرعية الصغيرة الأكثر قوة لتنبؤ هدف ما، وإبعاد الباقي.

- وتشمل منهجية ثانية، تمزج بعض متغيرات التنبؤ إلى مؤشرات، ومقاييس، ومعاملات (Factors)، وهي عملية تدعى استخراج السمة (Feature Extraction).

- أما المنهجية الثالثة، فتتجنب حيز قياس ضخم، من خلال إدراكها بأن حالات المزج الممكنة لقيم المتغيرات لن تكون ذات أهمية عملياً في المستقبل؛ إما لعدم وجود العديد من الحالات مع ذلك المزج الخاص، أو لأن المرء يمكن أن يحصل على تقديرات جيدة لتأثيرات المتغيرات الفردية دونما أخذ بعين الاعتبار كُلّ تفاعلاتها أو حالات مزجها الممكنة للقيم.

- تقسيم البيانات (Data Partitioning) (أو أشجار القرار) هو مثال من الأمثلة، وتبحث هذه الطرق عن التفاعلات الإحصائية بين المتغيرات، ولكنها لا تأخذ بعين الاعتبار بشكل شامل حيز القياس بأكمله مع ملايين تفاعلاتها أو خلاياها. إنها تشتغل - عوضاً عن ذلك - على متغير واحد تلو الآخر، باختيار، في الأول، المتغير الوحيد الذي يقسم البيانات بشكل أفضل، على كُلّ Y، ثم إيجاد - بشكل تكراري - متغيرات إضافية لتقسيم مزيداً من البيانات. إن هذه التقنيات تجد - دون شك - تفاعلات ذات أهمية

من حيث التنبؤ بـ Y، ولكن من خلال البدء بمتغير واحد تلو الآخر، سيحددون، ربما، مائة مزج مهم من قيم متغيرات (أو تفاعلات)، بدل مليارات. إن طرق الشجرة أو طرق التقسيم العودي، تنتقي مجموعة فرعية من المتنبئات والتفاعلات انطلاقاً من عدد أكبر من المتنبئات والتفاعلات الممكنة.

- الشبكة العصبية تلعب النماذج دوراً مماثلاً، إذ يمكن أن تدمج تفاعلات معقدة بين متنبئات ما في نماذجهم التنبؤية على نحو آلي، من دون أن يكون لدى محلل البيانات حاجة إلى تحديد تلك التفاعلات الأخرى التي أتت في وقت مبكر.

- **وأخيراً** تستفيد بعض الطرق من حقيقة إمكانية تقليص حيز قياس ضخم تقليصاً حاداً، إذا ما وضعنا افتراضاً مبسطاً، أي إن كُلّ متغير يؤثر في متغير تابع، بمعزل عن كُلّ متغير آخر، أو - على نحو أكثر دقة - إن المتنبئات مستقلة على نحو مشروط. ويعادل هذا، القول إن التفاعلات بين المتنبئات لا تهم. وتعد الطرق التي تلي هذا الافتراض المبسّط، بما في ذلك مصنف بايزن الساذج (Naive Bayesian Classifier) (الذي سيتم مناقشته في فصل لاحق)، دقيقة إلى حدّ ما في بعض السياقات.

ولنلخص فكرتنا بخصوص أن «البيانات الكبيرة، ليست كبيرة بما فيه الكفاية». لقد بدأنا بالإشارة إلى أن طرق الحضر في البيانات، يمكن - مبدئياً - أن تتبنى منهجية «شاملة» لاستكشاف بنية وتنبؤ، من خلال - مثلاً - دراسة كُلّ تفاعل ممكن بين المتنبئات، أو من خلال استعمال كُلّ متنبئ متاح. كما استخدمنا تجربة فكرة، لبيان عدم إمكانية استراتيجية شاملة في الغالب، باعتبارها أمراً عملياً، لأن عدد حالات المزج أو التفاعلات بين المتنبئات، تصبح كبيرة فلكياً، بل كبيرة جداً إلى درجة عدم حيازة أي مجموعة، بيانات حالات كافية لتناول كُلّ حالات المزج. ولما واجهت طرق التنقيب في البيانات ذلك، أصبحت تتبنى استراتيجيات بحث غير مستنزفة.

وهي لا تجرب عادة كُلّ الاحتمالات، على الرغم من أنها مازالت تجرب نماذج عديدة ممكنة. وعملياً، يقلص التنقيب في البيانات حيز القياس أو عدد الاحتمالات المقدرة. وهذه ناجزة بطرق شتى:

1. بواسطة اختيار ابتداء، مجموعة فرعية من متنبئات مهمة من قائمة أكبر - اختيار سمة.
2. من خلال مزج متغيرات داخل مقاييس (Scales)، أو مركبات (Composites) - استخراج سمة.
3. من خلال أحياناً، تجاهل تفاعلات بين المتنبئات للحصول على تنبؤ أكثر بساطة، ولكنه مع ذلك أكثر دقة.
4. من خلال البحث عن تجميعات لحالات مماثلة في البيانات، وتحليل كُلّ تجمع أو مجموعة بشكل منفصل.

الفصل الرابع

مراحل مهمة في مشروع التنقيب في البيانات

نظراً لما سبق أن قدمناه للبيانات الضخمة القليلة جداً، حيث أشرنا إلى التحديات التي تضعها بيانات عالية الأبعاد، يمكن الآن مناقشة الكيفية التي يتم بها الشروع في تحليل التنقيب في البيانات بشكل إجمالي. هناك ستة خطوات منفصلة من حيث التصور:

1. البتّ في إمكانية معاينة البيانات، وكيفية التعامل مع هذه المعاينة قبل تحليلها؛
2. بناء مجموعة غنية من السمات أو المتغيرات؛
3. اختيار السمة واستخلاص السمة؛
4. تشكيل نموذج ما أو عملية تناسبية باستخدام قائمة أصغر من السمات على بيانات التدريب؛
5. التثبت من ذلك النموذج أو إقراره من خلال بيانات الاختبار؛
6. تجربة طرق بديلة للتنقيب في البيانات، وربما مزج العديد منها (طرق مجموعة)، بغية استكشاف إمكانية تقديمها لحلّ أفضل. وفي هذا الفصل، نقدم مزيداً من التفصيل بشأن الخطوات الأربع الأولى.

متى تتم معاينة البيانات الضخمة؟

يتجنب علماء التنقيب في البيانات أحياناً، تحليل مجموعة بيانات كبيرة برمتها، إذ يعمدون عوضاً عن ذلك، إلى استنباط عينة عشوائية صغيرة من حالات منها، والشروع في تحليلها. وتُعزى إحدى أسباب القيام بذلك إلى كون الحاسوب السريع نفسه قد يعمل لساعات في تحليل ملايين الحالات، في حين إن التحليل المطابق الذي أُجري على عينة عشوائية، ومن ثم عينة تمثيلية لـ 20,000 حالة مثلاً، قد تكشف فقط عن الأنماط نفسها وتعمل بطريقة فائقة السرعة. وفي هذا المثال، تعد عملية معاينة البيانات الضخمة، مجرد طريقة من طرق تسريع التحليل، وتجنب إمكانية تحطم الحاسوب بسبب الذاكرة غير الكافية. (وقد تتمثل استراتيجية بديلة في إنجاز تحليل أولي باستخدام عينة عشوائية أصغر، وتحليل مجموعة البيانات برمتها في الآخر، بعدما يكون المرء قد بتّ في المتغيرات والنماذج القائمة على بيانات عينة أصغر).

وثمة سبب ثانٍ ومختلف جداً من وراء اقتطاف عينة من مجموعات بيانات كبيرة، يحدث عندما يكون باحث ما مهتماً بتنبؤ أحداث أو حالات نادرة نسبياً. ولهذا ربما يريد محلل ما - مثلاً - اكتشاف المعاملات الاحتيالية من خلال تحديد نمط مميز تشترك فيه تلك المعاملات. وقد تحتوي هذه القاعدة من البيانات لدى باحث ما، ملايين المعاملات الشرعية (مشفرة مثلًا بشفرة 0 بالنسبة إلى المتغير الهدف)، ولكن قد لا يحدد المعاملات الاحتيالية (مشفرة بشفرة 1) إلا ألفاً منها فحسب. وبتعبير آخر، هناك نسبة غير متوازنة (Lopsided) للغاية من الاحتيال في المعاملات الشرعية، ولكن هذه «إبر في كومة قش»، مهمة جداً.

وإن بعض تقنيات النمذجة والتصنيف لا تشتغل بشكل جيد مع التوزيعات غير المتوازنة المطبقة على المتغير التابع. وإن نموذج انحدار لوجيستي ما، مثلاً، الذي يواجه محصلات ذات تسع قيم صفر بالنسبة إلى كُلّ محصلة ذات قيمة صفر، يمكن أن يبني نموذجاً يتنبأ بالأصفار بشكل جيد جداً، ولكن على حساب فقدان العديد منها. قد يكون النموذج العام - نسبياً - مناسباً جداً بتصنيفه 95 في المائة من الحالات بشكل صحيح، ولكن مع ذلك قد يسيء تصنيف نصف حالات المعاملات الاحتيالية.

ولتجنب هذا النوع من المشاكل، من الأفضل - عندما يتم التركيز على محصلات

نادرة نسبياً - ضمّ كُلّ الحالات (مثلاً، الاحتيال) النادرة، واستنباط عينة عشوائياً، من النوع الآخر من الحالات الوافرة، للحصول على معدل قريب من 1:1 لهذين النوعين في مجموعة البيانات الجديدة. إن تقنيات التنقيب في البيانات ستقوم بتنبؤ أو تصنيف أفضل بكثير، بالنظر إلى وجود هذه المجموعة من البيانات المتوازنة نسبياً، مقارنة بإمكانية تطبيقها على عينة غير متوازنة للغاية. إن المعاينة قبل التحليل، أمر ضروري في هذا النوع من السياق.

بناء مجموعة غنية من السمات

قد يبدو من المفاجئ أن يبدأ علماء التنقيب في البيانات بشكل روتيني، مشروعاً ما من خلال تشكيل متغيرات جديدة، حتى لو سبق أن كان لمجموعة بياناتهم متغيرات أو سمات عديدة. في الواقع، يقضي بعض علماء التنقيب في البيانات وقتاً أكثر في تشكيل مجموعة غنية من السمات، مقارنة بما يقضونه في تشغيل النماذج. ويرجع ذلك إلى كون نجاح النمذجة، يقوم على امتلاك السمات الصحيحة، وأن الباحث قد لا يكون واثقاً مسبقاً بمتغيرات التنبؤ التي قد تكون تنبؤات أو مصنفات قوية. ومن الحكمة بداية أي مشروع من مشاريع التنقيب في البيانات انطلاقاً من تشكيل متغيرات جديدة، مع العلم أن قائمة المتغيرات يمكن تخفيضها بالتدريج لاحقاً، وترك فقط المتغيرات التي يتبين أنها متنبئات قوية لهدف ما أو لمتغير تابع، أو أنها تعمل بشكل جيد في تحليلات التجميع (Cluster Analyses)، أو أنها مصنفات (Classifiers).

وعملياً، يقوم علماء التنقيب في البيانات:

- باستشارة ما يسمى خبراء الميدان حول العوامل التي يحسبون أنها مهمة في المحصلات التنبؤية، ومن ثم تشكيل قياسات تمثل تلك العوامل. إن علماء التنقيب في البيانات، هم في الغالب غرباء دخلوا في منظمة لتحليل أنشطة سبق أن خَبَرَها المطلعون التنظيميون بشكل كبير. إنه لمن الحكمة استجواب هؤلاء الخبراء والحصول منهم على استبصاراتهم للاطّلاع على تشكيل المتغير.
- بخلق سمات جديدة، تعد نسباً (Ratios)، مشكلة من متغيرات قائمة؛ ففي

قطاع العقارات، مثلاً، قد يكون ثمن العقار للقدم المربع الواحد، قياساً أكثر فائدة من التكلفة الإجمالية أو الحجم الإجمالي لعقار ما. وفي البحوث الصحية، يعد مؤشر كتلة الجسم (Body Mass Index) نسبة معقدة، من الوزن إلى الارتفاع الذي يعمل بمثابة تنبؤ مفيد بالنسبة إلى أنواع مختلفة من المحصلات الصحية. وفي علم السكان، تُعد متغيرات حرجة عديدة معدلات (معدلات الأطفال بالنسبة إلى كُلّ 10,000 امرأة في سنّ الإنجاب، ومعدلات الطلاق بالنسبة إلى 1,000 زواج في السنة، وهكذا). ولهذا، فإن على علماء التنقيب في البيانات، ضمان التفكير في النسب والمعدلات المناسبة لدى تشكيلهم السمات في مجموعة بيانات ما.

- بتشكيل نسخ جديدة من متغيرات مستمرة، تهدف إلى ضبط التأثيرات اللا خطية لمتغير تابع هدف. ويمكن القيام بتوزيع الخانات (Binning) هذه، باستخدام تقسيم البيانات، أو برمجيات الشجرة، أو استخدام توزيع خانات مثالية (Optimal Binning)، كما تمت مناقشة ذلك آنفاً. وسيتم أيضاً تقديم أمثلة بهذا الشأن في فصل لاحق.

- بتشكيل متغيرات جديدة لتمثيل التفاعلات بين سمات أخرى أو متغيرات، ولكنها متغيرات يمكن أن تدخل - بعد ذلك - ضمن نماذج باعتبارها متغيرات في حدّ ذاتها. وسيحدد مربع لكشف عن التفاعل التلقائي (CHAID)، والتصنيف وشجرة الانحدار (CART) تفاعلات تم تفصيل القول فيها سلفاً.

- بالتذكير بأن بعض إجراءات التنقيب في البيانات تتطلب من الباحث إعادة قياس المتنبئات قبل تشغيل النماذج. ويتجلى مسوّغ إعادة قياس المتنبئات، في كون أن بعض المتغيرات تقاس بالوحدات مثل الدولارات، وتأخذ قيماً من صفر إلى مليون أو أكثر، في حين إن آخرين قد يكون لها فقط فئات قليلة (مثلاً، من واحد إلى خمسة)، وقد تبقى أخرى - مع ذلك - مجرد قيم عشرية تنحصر بين صفر وواحد. وقد تتحيز بعض تطبيقات التنقيب في البيانات لمتغيرات ذات مجموعة كبيرة من القيم أو تباين كبير، على حساب متغيرات

ذات مجموعة صغيرة من القيم. ويتمثل إيجاد حلّ لتلك المشكلة في إعادة قياس كُلّ المتنبئات، لتتساوى في نهاية المطاف. (تنجز بعض التطبيقات هذه العملية من عملية إعادة القياس تلقائياً، ومن ثم، لا حاجة للباحث لأن يقلق).

- بالتذكير بأن النوعين الأكثر شيوعاً من إعادة القياس، هما التقعيد (Standardization) (داخل درجات -z (Z–Score)) والمعايرة (Normalization). ويشمل التقعيد داخل درجة–z، تحولاً حسابياً: إن متوسط القيمة بالنسبة إلى ذلك المتغير، يُطرَح أولاً، من كُلّ قيمة مرصودة، والرقم المحصّل عليه، يُقَسّم بعد ذلك على الانحراف المعياري للمتغير. والتقنيتان كلاهما يعملان على جعل المتغيرات متساوية من حيث القياس.
- بالتذكير أيضاً بإمكانية إجراء تحليل تجميع ما، لتحديد المجموعات ذات الحالات المماثلة في مجموعة البيانات، دون الإشارة إلى المتغير التابع، أو المتغير الهدف. ويمكن للباحث اختيار عدد التجميعات مقدماً (غالباً حوالي أربعة). كما يمكن استخدام تلك التجميعات - إذن - لتحديد متغير عادي جديد، يمكن إضافته إلى مجموعة البيانات.

وباستخدام هذه الاستراتيجيات، ينتج علماء التنقيب في البيانات سمات أو متغيرات جديدة يتم توظيفها في مراحل متعاقبة في تحليل التنقيب في البيانات إلى جانب المتغيرات الموجودة سلفاً. وقد يتبين أن بعض المتغيرات الجديدة هي متنبئات مهمة، ولكن يمكن التخلص منها. وتعد هذه الغربلة دقيقة؛ ومن غير المرجح أن يهتدي المرء إلى نموذج قوي، ما لم يبدأ بمجموعة غنية من السمات.

وإن الأنشطة المتنوعة التي تنتج السمات، إلى جانب انتقاء المتنبئات الأكثر قوة (سيتم مناقشة ذلك في القسم التالي)، زائد البيانات المفقودة، كُلّ ذلك يشار إليه بالمعالجة المسبقة للبيانات (Preprocessing Data).

انتقاء سمة

إن طرق انتقاء السمة، تمكّن الباحث من تحديد المتَنَبِّئ المحتمل - من أصل

العديد من المتنبئات - المرتبط ارتباطاً قوياً بمتغير محصلة ما مهمة. وتساعد أيضاً على تجنب مشاكل ذات الخطية المتعددة (Multicollinearity) من بين متنبئات.

ويقدم التنقيب في البيانات بدائل متعددة، لانتقاء مجموعة فرعية من المتغيرات المستقلة التي تعد المُتَنَبِّئات الأكثر فاعلية لمتغير تابع. وهناك طريقة معروفة سلفاً لدى علماء الاجتماع الكميين: الانحدار التدريجي يعقبه منطقاً يشبه التنقيب في البيانات. وبعد تحديد متغير مستقل، يشتغل برنامج تدريجي موجه عبر كُلّ المتغيرات المستقلة، مُقَدِّراً لكُلّ منها قوة تنبؤية لنموذج انحدار، يضم فقط ذلك المتغير المستقل. ويختار المتنبئ الأفضل من هذه المتغيرات. وفي خطوة ثانية، يعود إلى مراجعة قائمة المتنبِّئات المتبقية، ويقيّم الأفضل منها الذي يحسّنُ من التطابق إذا ما أضيف إلى الأول في نموذج الانحدار. ويضيف ذلك المتنبئ الأفضل إلى المتنبئ الأسبق، ويكرر العملية مرات عديدة إلى غاية تحديد مجموعة فرعية من المتنبئات التي - إذا ما مزجت - تتنبأ جيداً بالمتغير التابع. إن طرق تقسيم الشجرة أو طرق التقسيم العَودي (Recursive Partitioning)، شبيهة بالانحدار التدريجي (Stepwise Regression) من حيث اختبارها كُلّ متنبئ محتمل على حدة وانتقاؤها المتنبئات الأكثر قوة، بينما تقوم أيضاً بتحديد تفاعلات بين متنبئات.

ويُزعم أن خوارزميات أخرى من خوارزميات التنقيب في البيانات لانتقاء السمة، تتفوق على الانحدار التدريجي، إما من حيث سرعة حوسبتها و/ أو من حيث كونها أقل تحيزاً. هناك مقاربة تعرف باسم الانحدار الأصغر للزاوية (LARS) أو الحدّ الأدنى للانكماش المطلق في أَمْثَلَةِ الانتقاء (Hastie, Tibshirani, and Friedman 2009; Miller 2002). ويتم تنفيذ انتقاء سمة اللاسو في الطبعة المهنية لبرمجيات إحصاء الحزمة الإحصائية للعلوم الاجتماعية (SPSS) (اختر الانحدار ← القياس الأمثل للانحدار القاطع، وانتق اختيار التسوية).

كما توجد خوارزمية سمة انتقاء أخرى، يُفترض أنها أكثر دقة من «اللاسو»، ولها أيضاً امتياز السرعة الفائقة، تدعى انحدار تباين عامل التضخم (VIP) (لأنها تستخدم عوامل تباين التضخم لانتقاء المتنبئات المحتملة)، وهي متاحة بالمجان بصفته برنامجاً في R (Lin, Foster, and Ungar 2011).

وداخل R، اكتب «Install:Packages» (VIF). وهناك معلومات إضافية على الرابط: http://cran.r-project.org/web/packages/VIF/VIF.pdf. وسنمثل لذلك في فصل لاحق.

استخراج سمة

تقدم ثلاث إجراءات، أدوات بديلة لبناء متغيرات جديدة، تعد مجموعاً مرجحاً (Weighted Sums) لمتغيرات قائمة، ويتعلق الأمر بتحليل المكوّن الرئيسي (Principal Components Analysis)، وتحليل المكوّن المستقل (Independent Component Analysis) ومزيج من الإسقاط العشوائي (Random Projection) وتجزئة القيمة المنفردة (Singular Value Decomposition).

ويعد تحليل المكوّن الرئيسي أكثر التقنيات رسوخاً وهو متاح في العديد من رزم البرمجيات، ولهذا فإننا ننصح باستخدامه. ويجد تحليل المكوّن الرئيسي مجموعة من المكوّنات (أو العوامل، أو المقاييس) التي - إن اجتمعت - ستفسر التباين الإجمالي داخل مجموعة بيانات بشكل أفضل (Dunteman 1989). ويتشكل كُلّ مكون من إضافة، عدد إلى متغيرات تم قياسها سابقاً، كُلّ بحسب عمله الترجيحي الخاص به. وتُنْتَقَى هذه المتغيرات وتُحسَب ترجيحاتها بطريقة تُفسِّر فيها المكوّنات أو المقاييس المستخلصة، التباين الشامل في مصفوفات البيانات قدر الإمكان.

وفي الخطوة الثانية، يتم «تَدْوير» هذه المكوّنات (Rotated) لتصبح حالياً غير مترابطة بأي بُعد آخر. وينتج عن ذلك عدد صغير من متغيرات جديدة أو سمات تلخص معظم التباينات الموجودة في العدد الأكبر من المتغيرات الأصلية. ومن ثم، فإن تحليل المكوّن الرئيسي يحقق تخفيض البعد من خلال تقليص عدد المتغيرات.

وليس أمر معطى، أن تكون المكوّنات التي تفسر قدراً كبيراً من التباين في مجموع بيانات، متنبئات جيدة لمتغير تابع خاص أو لهدف أو رقعة تعريف (Label). سيتم تحديد ذلك في مرحلة لاحقة خلال النمذجة. ومع ذلك، يخلق تحليل المكوّن الرئيسي متغيرات جديدة. ويمكن لبرمجيات تحليلية البتّ لاحقاً في المتنبئات الأفضل للنمذجة.

وتتمثل سلبيات تحليل المكوّن الرئيسي في كون المكوّنات أو العوامل التي تنتجها قد تفتقر إلى المعنى أو التأويل. و«يَحْمِلُ» إجراء تحليل المكوّن الرئيسي، متغيرات أصلية على مكونات، من خلال عملية ترجيح المتغيرات الأصلية بطريقة يفسر فيها المكوّن الكثير من التباين، ولكن من حيث التصور، لا يمزج ذلك - بشكل متكرر - متغيرات مختلفة جداً داخل مكون واحد. وماذا تعني إمكانية مزج مكون ما، لأسئلة أو لمقاييس حول أفراد مختلفين اختلافاً كاملاً، من خلال شحن مواقف تجاه الإجهاض بمقاييس دخل العائلة، والعمر، ومدة التنقل اليومي؟ وإذا تبين أن مكوناً مفككاً من حيث الموضوع، متنبئاً له دلالة في نموذج ما، فكيف يؤول المرء تلك الحقيقة؟

ويعيدنا هذا إلى التوتر القائم بين تحليل البيانات الذي يركز على آليات الفهم وعمليات سببية، مقابل تحليل يرتكز على دقة تنبؤية. وإذا كانت الغاية من وراء بناء نموذج ما، تتمثل في التنبؤ بدقة، وبعدها الاستناد في القرارات إلى ذلك التنبؤ، فإن عدم التماسك التصوري لمقاييس تحليل المكوّن الرئيسي، لا يهم كثيراً، مادامت «تشتغل». وإذا كانت الغاية، هي فهم عملية سببية ما، فسيطرح إنتاج تحليل المكوّن الرئيسي لمتنبئات غير متماسكة وغير قابلة للتأويل، مشكلاً.

إن برنامجاً لتحليل المكوّن الرئيسي، قد يعمل جيداً بمائة متغير، وبضعة آلاف حالة، ولكن تتباطأ وقد تنهار لدى مواجهتها مجموعات بيانات كبيرة جداً، لأن عملية معالجة مصفوفات ضخمة، أصبحت تستهلك - حسابياً - وقتاً طويلاً. ولكن لحسن الحظ أصبح لدى علماء التنقيب في البيانات القدرة على بسط منطق تحليل المكوّن الرئيسي ليشمل تحليل البيانات الكبرى ضمن مقادير معقولة لوقت المعالجة، من خلال مزج تقنيتين هما: الإسقاط العشوائي، وتجزئة القيمة المنفردة.

أولاً: يضرب الإسقاط العشوائي مصفوفة بيانات، في مصفوفة عشوائية، لإنتاج - في الواقع - العديد من المتغيرات الجديدة، بحيث يحتوي كُلّ واحد منها على متغير قديم، يُرجَّح بواسطة عدد عشوائي. وبعد ذلك، يضيف معاً، كُلّ تلك المتغيرات المرجَّحة حديثاً لإنتاج متغير جديد. وتبدو تلك، انطلاقاً من الانطباع الأول، فكرة غريبة جداً: إنتاج متغيرات جديدة، شبيهة بمقاييس، تعد مزيجاً عشوائياً بشكل دقيق

لمتغيرات قائمة من قبل. إن تجزئة القيمة المنفردة، إذن، تحلل هذه المتغيرات المنتجة حديثاً لخلق عدد أصغر من الأبعاد أو الخصائص التي يمكن استخدامها بعد ذلك، في نموذج من نماذج التنقيب في البيانات (Vempala 2004; Halko, Martinsson, and Tropp 2011). إن تجزئة القيمة المنفردة شبيهة بتحليل المكوّن الرئيسي بما أنها تقلص عدداً كبيراً من المتغيرات إلى متغيرات جديدة أقل.

لقد بين علماء الرياضيات إمكانية أن يحافظ العدد الأصغر لسمات أو متغيرات، تم إنتاجها بواسطة إضافة نسخ مثقلة عشوائياً من المتغيرات الأصلية، على البنية التي كانت موجودة في الأعداد الكبرى للمتغيرات الأصلية (Halko, Martinsson, and Tropp 2011; Martinsson, Rocklin, and Tygert 2011).

إنَّ تحليل المكوّن المستقل، مقاربة أخرى تطورت حديثاً لاستخراج سمات شبيهة بتحليل المكوّن الرئيسي. ويقول كتَّابُها بتفوقها - إلى أبعد الحدود - على تحليل المكوّن الرئيسي في قدرتها على إيجاد مكونات تتنبأ بهدف ما(Hyvarinen, Karhunen, and Oja 2001). هناك برنامج يدعى فاست آي سي أي (FastICA)، يمكن تحميله بالمجان من الموقع التالي: http://research.ics.aalto.fi/ica/fastica/.

إنشاء نموذج

بمجرد أن ينشأ باحث ما مجموعة بيانات، غنية من حيث السمات والمتغيرات، يمكن للنمذجة أن تبدأ. وسيختار عالم التنقيب في البيانات نوع النموذج المستخدم، ولكن هذه الخطوة الأولى، خطوة مرحلية فقط، بما أن باحثاً ما سيحلل البيانات، مستخدماً أنواع مختلفة عديدة من النماذج أو المقاربات، وسيقارن دقة تنبؤها قبل الاستقرار على مقاربة نهائية.

وإذا كانت غاية تحليل البيانات، هو التنبؤ بمتغير ثنائي (نعم/ لا)، فإن علماء التنقيب في البيانات لهم لوحة عريضة من المصنفات التي تنجز ذلك: طرق الجوار القريب (Nearest-Neighbor)، وطرق الشجرة، والمصنفات البايزية الساذجة (Naive Bayesian)، والمصنفات «البايزية»، وشعاع الدعم الآلي (Support Vector

(Machines)، والشبكات العصبية، إضافة إلى الطرق الإحصائية الراسخة القديمة، من قبيل الانحدار اللوجيستي، ووحدة الاحتمالية (Probit)، والتحليل التمييزي (Discriminant Analysis).

عندما يكون متغير الهدف أو المتغير التابع، قياساً مستمراً، فإن قائمة التقنيات القابلة للتطبيق تكون طويلة، بما فيها طرق الشجرة، ونماذج الشبكة العصبية، والانحدار البايزي (Bayesian Regression)، بالإضافة إلى مقاربات الانحدار التقليدي.

إن أي شخص حديث العهد بالتنقيب في البيانات، يسأل السؤال نفسه عند هذه النقطة: «ولكن ما هي الطريقة التي تعمل على نحو أفضل؟». الجواب عن هذا السؤال لن يروق لأحد: «يعتمد ذلك على كُلّ طريقة أو تقنية على حدة تم استعمالها»[1]. عندما حاول الباحثون مقارنة دقة هذه التقنيات المختلفة، مستخدمين مجموعات بيانات متعددة، لم يجدوا أي تقنية ما واحدة تتفوق على كُلّ التقنيات الأخرى بشكل متسق. وإذا حلل شخص ما مجموعة بيانات منفردة، فإنه في الغالب تتفوق تقنية على باقي التقنيات الأخرى، ولكن عندما ينتقل الشخص إلى تحليل مجموعة بيانات مختلفة، فإن ترتيب الطرق سيتغير بالكامل. والطريقة التي كانت بارزة من ذي قبل، هي الآن قريبة من الجزء الأسفل من القائمة، في حين ارتقت طريقة أخرى إلى أعلى القائمة.

ربما مع الوقت، سيطور باحثون نظرية، تمكن طرق التنقيب في البيانات من أن تكون الأنسب لمجموعات بيانات معينة، ولكن ذلك لم يحدث إلى حدّ الساعة. يبدو أن خصوصيات مجموعة بيانات ما، تهم حقاً - مظاهر بنيتها التي لم نستوعبها بسهولة. ولا يمثل ذلك حاجزاً عملياً أمام التحليل، وإنما يعني فقط أن أي عالم حساس من علماء التنقيب في البيانات، يجرب تقنيات نمذجة عديدة بالنسبة إلى مجموعة بيانات خاصة، ويلاحظ مدى أداء كُلّ تقنية ضمن هذا السياق الفريد بشكل جيد.

(1) لم يحسم المؤلف أمره بخصوص التقنيات المثلى التبعة في تحليل البيانات، لتوفر كُلّ تقنية على حدة، على ميزات خاصة بها (المترجم).

الجزء الثاني

أمثلة عملية

الفصل الخامس

إعداد التدريب ومجموعات بيانات الاختبار

منطق الصلاحية المتبادلة

ناقشنا سابقاً، كيف أن الصلاحية المتبادلة تعمل بمثابة آلية مراقبة جودة في عملية التنقيب في البيانات، وأشرنا إلى كيفية اختلاف طرق الصلاحية المتبادلة على نحو مهم، عن الاختبارات التقليدية من أجل دلالة إحصائية. سنناقش الآن بشكل واضح، منطق الصلاحية المتبادلة، وتقديم - بعد ذلك - دليل يبين كيفية تنفيذ هذه التقنية عملياً، مستخدمين عدداً من الحزم الإحصائية.

إن العديد من نصوص التنقيب في البيانات، تتناول منطق الصلاحية المتبادلة على نحو عابر جداً، بحيث يتم التركيز على تطبيقها العملي: كيف تقدّم الصلاحية المتبادلة حلاً لمشكلة يمكن أن تصادفه لدى استعمالها طرق مكثفة لمعالجة بيانات ضخمة. وفي بعض نصوص التنقيب في البيانات، تُقَدَّم الصلاحية المتبادلة باعتبارها طريقة تمنع التدريب المفرط (Overfitting) (Nibset, Elder, and Miner, 2009; Kuhn and Johnson 2013). وفي نصوص أخرى، تُستعمل باعتبارها طريقة لانتقاء نموذج ما (Murphy 2012)، كما تستعمل أيضاً في نصوص أخرى، وسيلةً من وسائل تقييم دقة النموذج (Han, Kamber, and Pei 2012). وفي الواقع، إن الصلاحية المتبادلة تشير إلى كُلّ ما ذكر عنها، ولكن لماذا هي على هذا النحو، ولماذا

تعد هذه المشاكل مترابطة، هما سؤالان غامضان. من أجل ذلك، نحاول هنا ملء هذه الثغرة التصورية.

إن الإشكالات التي يعالجها التنقيب في البيانات بالتوسل بطرق الصلاحية المتبادلة، هي إشكالات محورية ومألوفة في الممارسة العلمية. وفي البحث العلمي الذي يتناول إشكالاً بحثياً معيناً، لا يعتد أبداً بالنتائج الموجودة في دراسة مستقلة في حدّ ذاتها، باعتبارها وصفاً صالحاً لكيفية عمل العالم بشكل عام، لأن طبيعة الحظ للمعاينة أو أخذ العينات (Sampling)، وإمكانية الحدوث العشوائي في تجربة ما، تجعلها أمراً ممكناً، إلى درجة أن نتائج دراسة واحدة، هي نتيجة احتشاد ظروف عرضية. وكي يتم الوثوق بالنتائج والقبول بها، عليها أن تحصل على دعم من بحوث متعددة متتالية. وباختصار، لا بُدَّ من إعادة النتائج على نحو مستقل، ويكون ذلك مثالياً، إذا قام بذلك باحثون مختلفون كلياً.

تسمح إجراءات الصلاحية المتبادلة للباحثين باستعمال هذا المنطق داخل بحث أو تحقيق واحد. وفي طريقة الصلاحية المتبادلة الأكثر بساطة، يقسم الباحثون بياناتهم إلى عينات فرعية قبل بناء نموذج تنبؤي. ولأن تقسيم البيانات عشوائي، نظراً للغايات المحدودة للبحث المتوافر، فإن العينات الفرعية المولّدة تشكل مجموعات ترصدات مستقلة؛ فهي ليست مستقلة بمعنى شامل، بما أنها مستخلصة من الساكنة نفسها (أي المجموعة الكاملة من البيانات). ولكن داخل الكون الذي حُدّد من أجل الدراسة، وحُدّد أيضاً من قبل البيانات التي نحن بصدد استخدامها، تصبح المجموعات الفرعية مستقلة عن بعضها بعضاً عبر العشوائية (Randomization). ويبني باحثون نموذجاً، مستخدمين مجموعة من الترصدات وبعد ذلك، يقومون باختباره على نموذج آخر. وهذه الخطوة الأخيرة تمثل اختباراً مستقلاً لدقة النموذج.

تُستعمل العشوائية إجمالاً، لضمان أن المجموعات الفرعية - في المتوسط - التي تم إنتاجها، متشابهة قدر الإمكان من حيث الخصائص المناسبة (انظر مثلاً، Rubin 1978). ولكن لها نتيجة مريحة أخرى، تستغلها طرق الصلاحية المتبادلة للتنقيب في البيانات: التغير العشوائي بين المجموعات الفرعية التي تم إنتاجها عبر تعيين عشوائي (Random Assignment). وإن السمات التمييزية (Idiosyncratic

(Features) لأي مجموعة فرعية معينة تم خلقها بشكل عشوائي، قد لا تتكرر - أصلاً - في المجموعات الفرعية الأخرى. وفي المقابل، من المرجح أن تكون الانتظامات التجريبية عبر المجموعات الفرعية خصائص الساكنة بأكملها، أي إنها ستخبرنا بالإشارة (Signal) التي نريد الانتباه إليها؛ إن التغير العشوائي عبر العينات الفرعية يمكن التفكير فيه باعتباره الضجيج الذي نود أن نفصله تحليلياً.

ولهذه أهمية خاصة لمرونة طرق التنقيب في البيانات وقوتها الشديدة في إنتاج نماذج تنبؤية. ولأن نماذج من قبيل الشبكة العصبية، وأشجار التقسيم قادرة على مطابقة نفسها بشكل وثيق مع البيانات، فهي عرضة للسماح للضجيج بالقيام بدور أكبر مما يرغب فيه المرء في توليد النموذج الذي يقوم على أي مجموعة خاصة من مجموعات الترصدات. وتصبح لدانة (Plasticity) النماذج وقوتها - بهذا المعنى - هي لعنة. إذ سينتجون نموذجاً دقيقاً للغاية، محققين كمالاً تقريباً، في الدقة التنبؤية بالنسبة إلى المجموعة الخاصة من الترصدات التي قامت عليها. ولكن هذه النتيجة مخيبة للآمال، أو «مفرطة في التفاؤل» بتعبير أحد الباحثين القدامى (Larson 1931)، بما أن هذا النموذج لن يكون أداؤه جيداً أبداً انطلاقاً من عينة. وهذا ما يشير إليه متخصصو التنقيب في البيانات بالإفراط في التدريب (Overfitting). ويمكن أن يساعد استخدام الصلاحية المتبادلة المساعدة في انتقاء النموذج وتقييمه، وتقليص إمكانية حدوثه.

لنتأمل سبب أهمية هذا في نهاية المطاف. إن خوارزميات التنقيب في البيانات، يمكن أن تكون أدوات قوية للتنبؤ، كما يمكنها - من ثم - تحسين القدرة العملية التشخيصية؛ أي إنه، إذا تدربت خوارزمية من خوارزميات التنقيب في البيانات على مجموعة من البيانات التي يستطيع الباحثون من خلالها الولوج إلى القيمة الحقيقية لمتغير النتيجة (Outcome Variable)، والتحقق من مجموعة مستقلة حيث قيم النتيجة معروفة أيضاً، يمكن نشرها لاحقاً في بيانات حيث متغير النتيجة - أي كمية الفائدة - غير معروف. على سبيل المثال، يمكن لأدوات التنقيب في البيانات تحسين قدرة الممارسين على التمييز بين الخلايا السرطانية، وغير السرطانية. وبالنظر إلى هذا التطبيق العملي، فإنه بات من الأهمية بمكان، أن تكون النماذج دقيقة، من خلال تطبيق صارم للصلاحية المتبادلة من أجل انتقاء نموذج وتقييمه.

لدينا تعقيب أخير حول أهمية طرق الصلاحية المتبادلة. وقد بات لدينا يقين في مسار بحثنا أن القوة التنبؤية لطرق التنقيب في البيانات الكثيفة حسابياً (Computationally Intensive) تأتي بتكلفة، من حيث قدرة النماذج على فهمها فهماً تاماً من قبل البشر. ويصف كوهن (Kuhn) وجونسون (Johnson) (2013) هذا «بالتوتر بين التنبؤ والتأويل». وغالباً ما تدخر نماذج التنقيب في البيانات قوتها عبر تعقيد متزايد، مما يجعلها مربكة، وإن لم نقل ببساطة إنها مبهمة بالنسبة إلى المحللين البشر. ولكن يرى كوهن وجونسون - خاصة في حالات الحياة والممات - عدم إيلائها أي أهمية، وأن تفضيل نموذج مفهوم وذو أداء ضعيف نسبياً، على حساب نموذج ذو «علبة سوداء» وتنبؤي للغاية، هو أمر «غير أخلاقي». وفي سياق النماذج التي تعجز نتائجها عن أن تكون مفهومة بسرعة، ويوحي ظاهرها بأنها دقيقة للغاية، تبدو الصلاحية المتبادلة وسيلة أساسية من وسائل توليد الثقة عبر الاختبار الصارم.

وباختصار، تمنح الصلاحية المتبادلة اختباراً مستقلاً للنموذج المتطور بتقنية من تقنيات التنقيب في البيانات. فهي تساعد على اختيار النموذج «الأفضل» (انتقاء النموذج) من حيث قدرته على التنبؤ من عينة، وعلى تقييم القدرة التنبؤية «الحقيقية» لنموذج ما (تقييم النموذج). ويساعد هذا على الاحتراز من إمكانية انتقاء نموذج متوقف أساساً على البيانات الخاصة التي كانت تقوم عليها - أي إنها تحترز من التدريب المفرط. كُلّ هذه الوظائف مترابطة بإشارة مشتركة إلى منطق الاختبار المستقل وموثوقية النتائج القابلة للتكرار. والآن نعرض لنقاش مختصر للطرق المختلفة للصلاحية المتبادلة، وبعد ذلك نشرع في توضيح كيفية أداء الصلاحية المتبادلة، من خلال التوسل بعدد من البرامج الإحصائية.

طرق الصلاحية المتبادلة: نظرة شاملة

يمكن توليد مجموعات البيانات «المستقلة» بطرق شتى، غير أن الطريقة الأكثر بساطة من حيث التصور، والأكثر تعقيداً - في الوقت نفسه - من حيث العمل الحقيقي المطلوب، تتمثل في الاشتغال ببيانات جُمعت على نحو منفصل. وإذا قمنا ببناء نموذج تنبؤي لمعدل الوفيات، متوسلين ببيانات مأخوذة من مستشفى واحد، فسيكون بإمكاننا إخضاعه للاختبار انطلاقاً من بيانات تم جمعها في مستشفى

مختلف، ولكن هذه الحالة نادرة إلى حدّ ما، على الرغم من أنها مرغوب فيها، على ما يبدو. إن عملية جمع البيانات مكلفة، ومن غير المرجح أن يرغب الباحثون والممولون في مضاعفة تكاليف البحث لمجرد قدرة المُنَمذِجين التنبؤيين الحصول على مجموعة بيانات اختبار نظيفة.

ومن ناحية أخرى، توجد ثلاث طرق يتم من خلالها توليد البيانات المستقلة من مجموعة بيانات وحيدة: النظام التمهيدي (Bootsrapping)، والكابح العشوائي (Random Holdback)، وطية-ك (K-Fold). وتشمل الطريقة الأولى، المعاينة العشوائية بالاستبدال، من أصل البيانات المتوافرة لدينا. وغالباً ما يتم القيام بهذه العملية، مرات عديدة؛ فتنشأ مجموعات بيانات عديدة منفصلة بحجم مساوٍ لحجم مجموعة بياناتنا الأصلية. وإذا كنا نظن أن بياناتنا الأصلية، كانت عينة عشوائية نسبياً، مأخوذة من الساكنة، فإن «النظام التمهيدي» يقدم طريقة غير متحيزة لتوليد عينة عشوائية لكُلّ العينات العشوائية الممكنة. كما يمكن استخدام كُلّ عينة من العينات الممهَّدة (Bootstrapped) المتعددة في تحليل البيانات، وتقديم مجموعة من النتائج (المعاملات أو الاحتمالات المتنبأة، على سبيل المثال)، التي يمكن بعد ذلك إيجاد متوسط لها للحصول على نتيجة عامة.

ويقدم «النظام التمهيدي» بعض الجوانب الإيجابية المتفوقة على طرق أخرى، سيتم مناقشتها لاحقاً. كما يمكن بخاصة - من خلال البوتسرابينغ - توليد تقدير لمعدل الخطأ، إلى جانب توليد خطأ معياري للتنبؤ، وكذا خطأ معياري لمعدل الخطأ على حدٍ سواء[2] (Efron 1979, 1983; Efron and Gong 1983). علاوة على ذلك، ينتج «النظام التمهيدي» خطأً معيارياً لا معلمياً (Nonparametric) - لا يقوم على افتراضات توزيعية قد تفتقد إلى السند التجريبي - دخل في واقع الأمر، في

(2) إن معدل الخطأ مساوٍ لعدد الحالات التي أساء النموذج تقسيمها (Misclassified)، مقسوم على مجموع عدد الحالات. ومن خلال البوتسرابينغ، يمكننا توليد توزيع معدل خطأ، الذي يعد عادياً على وجه التقريب. وسيكون متوسط التوزيع عادة أكبر من التقدير الساذج لمعدل الخطأ، انطلاقاً من بيانات التدريب، وسيشكل تقدير «الواتسراب» لمعدل الخطأ. ومن الممكن أيضاً حساب خطأ «الواتسراب» المعياري الذي يعادل الانحراف المعياري لمعدلات الخطأ في مجموعة العينات الممهَّدة التي تم إعادة تشكيلها، مقسوم على جذر مربع لحجم العينة. (المترجم)

النمذجة الإحصائية على هذا الاعتبار أساساً. وأخيراً، إن «النظام التمهيدي» ينتج تقديراً ممهداً (Estimate Smoothed) لمعدل الخطأ، لأنه يولّد مجموعات بيانات متعددة، عوض مجموعة بيانات واحدة، التي على أساسها يتم اختبار النموذج. ومع ذلك، إن «النظام التمهيدي» كثيف جداً حسابياً، وقد يكون مستنزِفاً للوقت لدى استخدامه مع مجموعات بيانات كبيرة.

وتعد مقاربة «النظام التمهيدي» قيّمة بخاصة عندما يكون لدى المرء عينة صغيرة، يبدأ بها عمله، حيث القوة الإحصائية تشكل فعلاً قضية من القضايا. ولكن في حالات أخرى، تكون لدينا بيانات تحتاج إلى الصقل، وهو أمر يزداد صحة في عصر البيانات الضخمة. وفي هذه الحالة الأخيرة، يمكننا تبني مقاربة أكثر بساطة للغاية من البوتسرابينغ، تدعى الكابح العشوائي (Random Holdback): يمكننا ببساطة تقسيم البيانات إلى مجموعة تدريب ومجموعة اختبار، وبناء نموذجنا باستخدام الأول، وإخضاع الثاني للاختبار. كما يمكننا تقسيم البيانات بين أجزاء التدريب وبين أجزاء الاختبار إلى ما نشاء من النسب – 50/ 50، 70/ 70، وهكذا.

وفي المقابل، يمكننا تبني مقاربة ثالثة باستخدام ما يُسمى بالصلاحية المتبادلة لطية-ك (k-fold). ويشمل هذا، تقسيم البيانات بشكل عشوائي إلى أجزاء-ك (أو طيات-ك)، ذات حجم متكافئ، بحيث تكون (ك) عددا من اختيار الباحث. (وأما القيم النموذجية لـ (ك)، المستخدمة في حزم البرامج الشعبية، فهي 5 أو 10 طيات)؛ ثم يبني الباحث نماذج (ك) منفصلة، بحيث يستخدم كُلّ واحد منها فقط طية من الطيات. ويتم اختبارها بعد ذلك، على آخر ما تبقى من الطيات. وفي جوهر الأمر، لدينا مجموعة تدريب (ك)، ومجموعة اختبار (ك)، وكل نموذج من نماذج (ك)، المحصل عليها، يتم اختبارها على بيانات لم تستخدم في توليدها. ثم يمكن ادماج النتائج من خلال إيجاد المعدل، أو يمكن اختبار النموذج المناسب جداً. وتعد هذه الطريقة أيضاً جيدة بالنسبة إلى مجموعات بيانات صغيرة نسبياً.

وسنبين الآن كيفية إنجاز طية-ك والصلاحية المتبادلة الكابحة باستخدام حزم برمجية إحصائية مختلفة.

«الستاتا»

إن «الستاتا» ليست في الحقيقية حزمة من حزم التنقيب في البيانات، وعليه فإن الصلاحية المتبادلة لم تبنى داخلها على نحو يجعلها سهلة الاستخدام بشكل خاص. ومن الضروري - على حدّ علمنا - القيام بالصلاحية المتبادلة الكابحة «باستخدام اليد».

«إن الأمر الأول الذي يجب القيام به لإنجاز الصلاحية المتبادلة الكابحة، هو تقسيم البيانات بشكل عشوائي إلى جزئين». وتسمح «الستاتا» بمعاينة عشوائية، تمكن الباحث بعدها، من إنتاج مجموعة بيانات منفصلة، غير أننا نظن أنه من اليسير جداً التوليد ببساطة، متغير يسمح بتقسيم عشوائي. ويمكن القيام بذلك من خلال توليد معادلة ذات حدين (Binomial) باستخدام الشفرة التالية:

```
gen x = rbinomial (n,p)
```

الجدول رقم 1.5: نتائج مأخوذة من برنامج الطية المتبادلة (Crossfold) لـ «الستاتا».

عدد الطيات المتبادلة	خطأ جذر متوسط المربعات
1	0.3352256
2	0.3308182
3	0.3365854
4	0.3280907
5	0.3264875

وتعمل هذه التسمية على توليد متغير جديد يدعى *x*، بحيث يوَزع باعتباره معادلة ذات حدين. ويمثل المعلم *n*، عدد تجارب المعادلة ذات حدين تبعاً لكُلّ حالة، في حين تمثل *p* احتمال «النجاحات». أما بالنسبة إلى الإسناد العشوائي، فتقوم $n = 1$ و*p* على انهيار التدريب/ الاختبار الذي ترغب فيه. وإن تحديد *p* في قيمة .5، سيولّد متغيراً جديداً ذا انهيار يقدر بـ 50٪ - 50٪، للوحدات والأصفار؛ كما أن تحديد *p* في قيمة .7، سينتج انهياراً بواقع 70٪ - 30٪، وهكذا.

وأما الاختبار المستقل، فتلك قصة أخرى. ففي حدود علمنا، قد يشمل هذا بناء نموذج (انحدار لوجيستي) انطلاقاً من قسم من البيانات (حيث x = = 1)، بحيث تُخزّن تقديرات المَعلَم في المتجه (Vector)، ثم يتم تنبؤ الحصيلة في باقي البيانات (حيث x = = 0)، وذلك باستخدام معادلة انحدار مولّدة عبر ضرب مصفوفة متغيرات في هذه المتجهة من المعامِلات. وتعد هذه العملية شاقة نوعاً ما، ويدخل المستعمل في حقل برمجة مصفوفة في «الستاتا»، التي تعد مقدمة إلى حدّ ما. وباختصار، يصعب القيام بتثبيت الكابح (Holdback) في «الستاتا»، لأنَّ مصمميه لم يبنوا – من حيث المبدأ – النظام وفي ذهنهم انشغالات تتصل بالتنقيب في البيانات.

ويمكن القيام بالصلاحية المتبادلة لطية-ك مباشرة بواسطة برنامج يولّده المستخدم، يُدعى الطية المتبادلة (Crossfold)، (Daniels 2012). وتستخدم الطية المتبادلة صياغة (Syntax) «الستاتا» التالية:

Crossfold regress yvars xvars, k (k)

ويمكن استبدال «الانحدار» في هذه الصياغة باللوغاريثم (Logit) أو مقدِّرات (Estimators) أخرى. ولسنا واثقين – التوثيق لا يخبرنا بأي شيء – على وجه التحديد، من عدد المقدّرات التي تدعمها الطية المتبادلة. وعلى كُلّ حال، فهي تقدم إحصائيات تطابقية انطلاقاً من نماذج ك (k) (مع اختيار الباحث لـ ك (k). وتسمح باختيار الإحصائيات التطابقية – خطأ جذر متوسط المربعات (RMSE)، خطأ المتوسط المطلق، أو R^2 – الزائفة. وثمة نتيجة نموذجية، يبينها الجدول رقم 1.5، غير أنها لا تمثل تقنية الصلاحية المتبادلة الأكثر إفادة، بل تقدم اختبارات ك (k) المستقلة لنموذج ما.

R

لسنا على دراية بروتين R المعين، الذي ينتج صلاحية متبادلة كابحة (Holdback CV)، وقابل للمزج بأي روتين تحليلي. هناك بعض روتينات R، تدمج الصلاحية المتبادلة عبر الكابح (مع افتراض أن المرء سبق أن أنتج مجموعات اختبار وتداريب).

وتعد الصلاحية المتبادلة الكابحة بالنسبة إلى هذه الروتينات سهلة للغاية، في حين تعد صعبة بالنسبة إلى آخرين، صعوبة وجودها في «الستاتا».

ومع ذلك، توجد روتينات الصلاحية المتبادلة لطية-ك في R. وإن cv.glm واحدة من روتينات الصلاحية المتبادلة لطية-ك، التي تعد حدسية بخاصة. وهي جزء من الحزمة التي يطلق عليها اسم بوت (Canty and Ripley 2010). وتستخدم في عملية صلاحية متبادلة تهم نماذج خطية عامة تطابقية سابقة. ويمكننا تسمية ذلك باستخدام الصيغة التالية:

```
cvl<-cv.glm (data, glmfit, k)
```

وحيثما كانت البيانات (Data)، تمثل مجموعة بيانات (Dataset) ما، إلا ومثلت glmfit النتائج لتطابق خطي عام سابق للنموذج، من خلال استخدام البيانات، و(k) هو عدد الطيات المرغوب فيها. وبعد تشغيل هذا النموذج، وإدخالـ

```
cvl$delta
```

تعود متجهة العددين: تطابق صلاحية متبادلة، وتطابق صلاحية متبادلة معدلة (إذا أُدخلت قيمة لـ (k)، عوض إسقاطها من الصلاحية المتبادلة المفترضة)[3]. ومُنح العدد الأخير لأن عملاً إحصائياً ما يقترح إمكانية أن يولّد إسقاطه قيمة لـ (k) من الصلاحية المتبادلة، تقديرات متحيزة لتطابقية الصلاحية (Davidson and Hinkley 1997). من أجل هذا، ينجز البرنامج بعض العمليات للتعويض عن هذا التحيز.

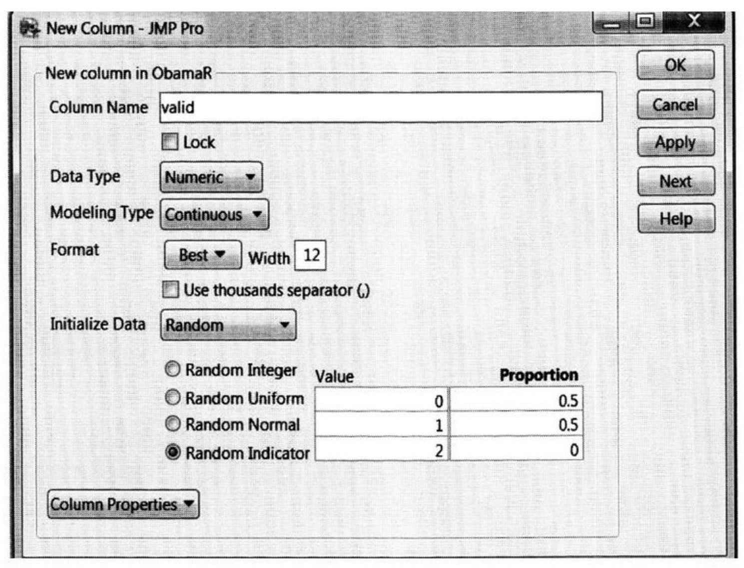
«غامب برو»

تعد الصلاحية المتبادلة سهلة للغاية في «غامب برو»، وتتم بطريقتين. فبالنسبة إلى بعض الروتينات المُنَمذِجة، يتطلب «الغامب برو» توفير متغير صلاحية - أي

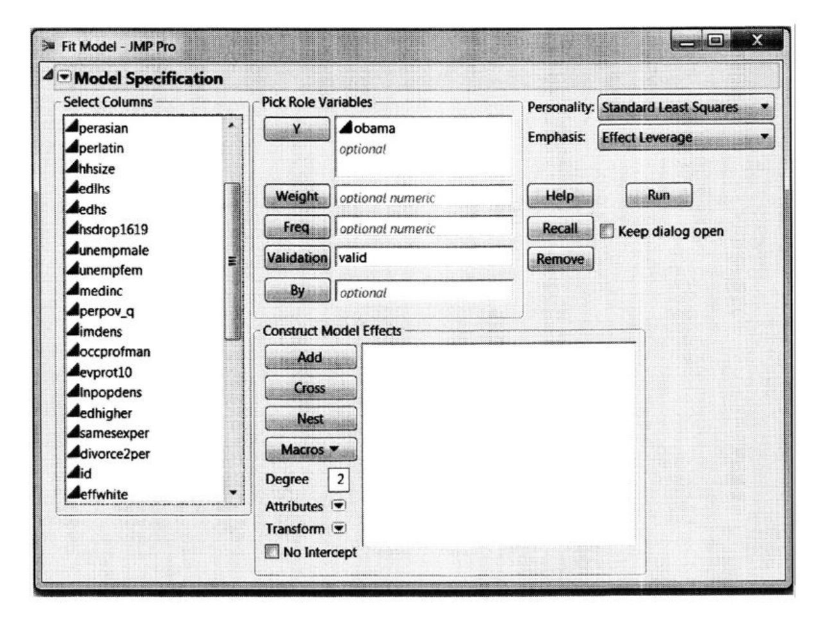
(3) وفي حالة إسقاط قيمة لـ ك (k) من الصلاحية المتبادلة، بالنسبة إلى مجموعة بيانات تضم ترصدات n، يُبنى نموذج ما على ترصدات n-1، وبعد ذلك، يتم اختباره على الترصد المتبقي. ويتكرر هذا n مرات. وفي الحقيقة، يعد إسقاط قيمة لـ ك (k) من الصلاحية المتبادلة حالة خاصة من الصلاحية المتبادلة لطية-ك، حيث ك (k) مساوٍ لـ n. (المترجم)

متغير اسمي (Nominal Variable)، يضم قيماً مختلفة تشير إلى أقسام التدريب والاختبار. وتدعو الحاجة إلى متغير يتم فيه تعيين القيم بشكل عشوائي في النسب التي نرغب فيها، على أن يتم ذلك بسهولة عالية. وفي القائمة الرئيسة «للغامب برو»، انقر كولز نيو كولم (Cols New Column). أما الويندوز (Windows)، الذي يُفتح، فهو مبين في الشكل رقم 1.5.

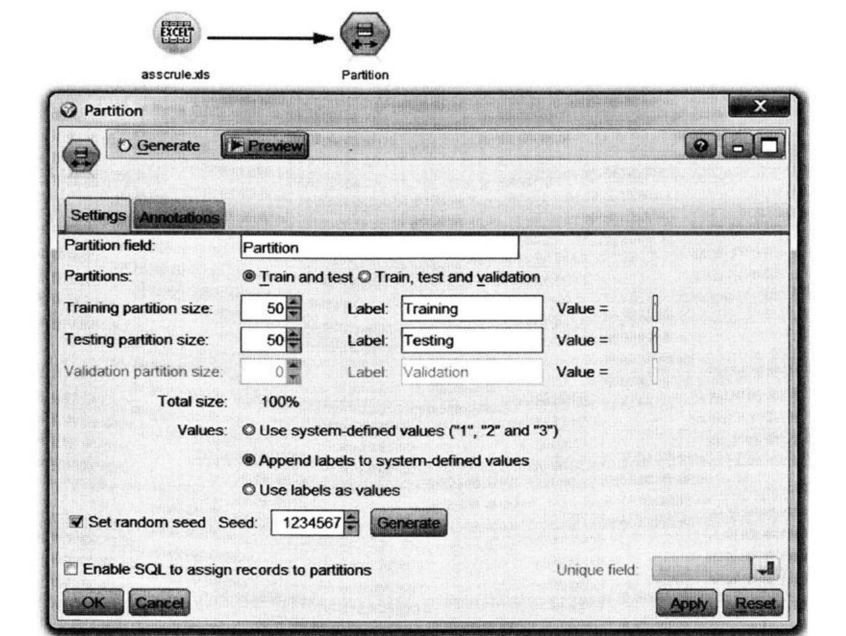
ونغير اسم العمود (Column) من القيمة الفرضية (Default) إلى قيمة مضبوطة أو صالحة (Valid) (كما هو الحال في الصلاحية (Validation)). ونغير النوع المنمذِج (Modeling) من القيمة الفرضية المستمرة (Continous) إلى اسمية (Nominal)، وبعدها ننتقي عشوائية تحت قائمة مهّد البيانات (Initialize Data). ويمكننا ذلك من اختيار توزيعات مختلفة، تُسحب من: العادي، والمتجانس أو المنتظم (Uniform)، والثنائي ذي الحدين (وهو ما يسحبه الدال العشوائي (Random Indicator)). وننتقي الدال العشوائي، ونغير النّسب للقيمتين 0، و1 إلى 0.5 و 0.5 (وتحدد قيمة 2 في 0). وببساطة، فإن نقر OK، يستمر في توليد متغير الصلاحية. كما يمكن ببساطة إدخال المتغير في مجالات الصلاحية «ويندوزات» لاحقة من «ويندوزات» بناء النماذج، كما هو الحال في البرنامج (Platform) المنفَّذ للانحدار التدريجي (Stepwise Regression)، والمبين في الشكل رقم 2.5. كما تملك برامج (Platforms) أخرى لنموذج ما مواقع صلاحية متبادلة مدمجة (Built-In) في الغامب (JMP). وعلى سبيل المثال، يمكِّن برنامج التقسيم (Partition) (بالنسبة إلى أشجار التقسيم) الباحث من الإشارة إلى «قسم الصلاحية» (Validation Portion) في «ويندوز» برنامج النموذج الأساسي. كما يمكّن المستخدمين من اختيار الصلاحية المتبادلة لطية-ك في «ويندوز» المنفَّذ للنموذج. وتمكن الشبكات العصبية كذلك المستخدمين من تحديد قسم كابح. وسنعرض إلى ذلك بتفصيل أكبر في شجرة التقسيم وأقسام الشبكة العصبية أدناه.



الشكل رقم 1.5: إنتاج عمود صلاحية في «غامب برو» (JMP Pro.).



الشكل رقم 2.5: إضافة الصلاحية المتبادلة للانحدار التدريجي في «غامب برو».



الشكل رقم 3.5: الصلاحية المتبادلة في نموذج الحزمة الإحصائية للعلوم الاجتماعية (SPSS).

نموذِج الحزمة الإحصائية للعلوم الاجتماعية

إن نَموذَج الحزمة الإحصائية للعلوم الاجتماعية، أي حُزمة محلل البيانات المتخصصة في إحصائية العلوم الاجتماعية يسهّل الصلاحية المتبادلة بخاصة، عن طريق الكابح (Holdback). وإن البرنامج الذي سنصفه لاحقاً بتفصيل أكثر، يشمل توليد تدفقات (Streams) عمليات إحصائية عبر الإشارة والنقر (Point-and-Click)؛ بحيث يضم كُلّ تدفق «عُقَداً» (Nodes)، قادرة على إنجاز عمليات، ولكل عقدة «ويندوز» مرتبط، يمكن من خلاله مواءمة معلمات متعددة. ويتم انتقاء العُقد من «لوحات الألوان» التي تضم عقداً مماثلة.

وفي مجالات لوحات الألوان، اختر عقدة التقسيم، وانقر مرتين كي تفتح الشاشة المبينة أعلاه في الشكل رقم 3.5. وداخل العقدة - كما يمكن مشاهدة ذلك - يمكن

للمرء تعيين انهيار عينات التدريب والاختبار (والصلاحية). ويولّد ذلك متغيراً يُدعى التقسيم (أو اسم آخر إذا ما غيّره الباحث)، الذي يمكن انتقاؤه بصفته متغير صلاحية في نمذجة العُقد.

ويمكن إنجاز طية-ك أيضاً لفائدة بعض التطبيقات الأخرى في النموذج أو «المودلير» (Modeler) (أقرب الجيران لـ «ك»، شجرة التقسيم 0.5، الشبكة العصبية)، ولكن داخل العُقد بالنسبة إلى هذه العمليات النموذجية المحددة، وليس باعتبارها عقدة منفصلة.

وظلت آي بي إم تحسن من إحصائيات الحزمة الإحصائية للعلوم الاجتماعية (SPSS) - وهو برنامج الإحصائيات المنتظمة المستخدمة في مئات الفصول الدراسية الجامعية - من خلال تطبيقات التنقيب في البيانات المتعددة. وإن لبعض من هذه التطبيقات خيارات داخلية بالنسبة إلى الصلاحية المتبادلة. ومع ذلك، من السهل - بما يكفي - تقسيم أي مجموعة بيانات من بيانات الحزمة الإحصائية للعلوم الاجتماعية إلى قسمين عشوائيين - بالنسبة إلى التدريب والاختبار على حدّ سواء - باستخدام الصياغة العادية للحزمة الإحصائية للعلوم الاجتماعية. وفي المثال أدناه، قمنا بتقسيم بياناتنا عشوائياً، بحيث حُدّدت 80٪ من الحالات، باعتبارها تدريباً، وحدد ما تبقى (20٪) باعتباره اختباراً؛ هذا، وبإمكان المستخدمين اختيار نسبهم الخاصة بهم. ويمكن قراءة الصياغة على النحو التالي:

```
USE ALL.

COMPUTE filter_$ = (uniform (1) < = .80).

VARIABLE LABELS filter_$ 'Approximately 80 % of the cases (SAMPLE)'.

FORMATS filter_$ (f1.0).

FILTER BY filter_$.

EXECUTE.
```

```
Filter OFF.
Recode filter_$ (0 = 0) (1 = 1) into datagroup.
var label datagroup ‘training or test’.
value labels datagroup 0 ‘test dataset’ 1 ‘training dataset’.
execute.
```

وتأخذ مجموعة البيانات (Datagroup) المتغيرة قيم 1 بالنسبة إلى الترصدات التي تم إسنادها بشكل عشوائي إلى 80٪ من فرعية التدريب، 0 بالنسبة إلى تلك المستخدمة في فرعية الاختبار.

الفصل السادس
أدوات انتقاء المتغير

عندما نُحلل بيانات ضخمة، نواجه سيلاً من المعلومات، ولدينا حالات عديدة، أو معلومات كثيرة عن كُلّ حالة من أجل استخدام فعال لمناهج إحصائية معيارية. وسبق لنا أن رأينا كيف أن مسألة امتلاك حالات كثيرة جداً، يمكن أن يتسبب في توقف البرامج أو في اشتغالها ببطء على نحو غير ملائم، كما رأينا كيف يمكن لهذا - أحياناً - أن يتناول ببساطة عن طريق معاينة بياناتنا. وتظهر حالة أكثر تعقيداً عندما تكون لدينا معلومات كثيرة جداً عن كُلّ حالة، وبتعبير آخر عندما تكون لدينا متغيرات أكثر مما ندرك ما نقوم بها.

ويستخدم مختصون في التنقيب في البيانات، حرف N للإشارة إلى عدد الترصدات أو الحالات، وحرف P للإشارة إلى عدد المتغيرات، أو المتنبئات، أو السمات. وفي الحالة التي تكون فيها P كبيرة جداً، نبحث عن تقنية تقلص مقدار المعلومات التي نحتاج في المعالجة من خلال انتقاء تلك المتغيرات ذات الأهمية القصوى والتخلص من الآخرين.

ثمة حلّ للتحول إلى مجموعة من التقنيات، تدعى طرق انتقاء مجموعة فرعية (Subset Selection)، أو طرق الضبط. وقد تم تطوير هذه النوع من الطرق - تحديداً - من أجل أتمتة عملية انتقاء المتغير، (ولهذا السبب نفسه، فهي غالباً ما تنتقد من قبل

محللي بيانات تقليديين لكونها غير نظرية (Atheoretical)، ومقيدة بالبيانات)، وتعمل الطرق قيد البحث، من خلال استكشاف - من بين قائمة طويلة من المتنبئات - تلك التي تؤدي أداء جيداً من حيث شرح التباين على مستوى المتغير التابع.

وقد كان الانحدار التدريجي الروتين «المؤتمت» الأول الذي تم تطويره لاختيار المتغيرات، وهو لا محالة، الانحدار الذي لقي استهجاناً شديداً - بشكل متكرر - من قبل الرافضين لهذه الخوارزميات. وتم تطوير الانحدار التدريجي، بالنسبة إلى حالات رُزِقَ فيها باحث ما، بوفرة المتغيرات المستقلة في مجموعة بيانات (أي P كبيرة)، لكن لديه إطار نظري محدود أو منعدم لاختيار الأنسب منها لضمه إلى نموذج ما. ومن المحتمل أن تبقى الحالة نفسها التي تستعمل فيها في معظم الأحيان، ولكن هناك حالات أخرى يمكن أن تستعمل فيها بشكل مثمر، ويكون لاستعمالها - على ما يبدو - أكثر من مبرر.

وفي تجربتنا، يمكن استخدام الانحدار التدريجي ليس فقط للتخلص من أعداد هائلة من متنبئات محتملة، تمثل التأثيرات الرئيسة، بل أيضاً للتدقيق في الشروط ذات الترتيب التفاعلي العالي بين المتنبئات. لنتصور أن لدينا اثنتا عشر متغير تنبئي، نود ضمها إلى نموذج ما. ولكن نريد أن نكون واقعيين بشأن حقيقة أن العالم لا يضم فقط التأثيرات الرئيسة، وإنما أيضاً التفاعلات بين المتنبئات، ونريد أن نأخذ بعين الاعتبار إمكانية تفسير بعض التفاعلات، تبايناً جوهرياً في متغير النتيجة (Outcome Variable). وإذا ما رغبنا في ضمّ تفاعلات في اتّجاهين، فإن عدد السمات أو المتنبئات في النموذج ترتفع من 12 إلى 88. كما يرتفع هذا العدد إلى 100 إذا ما عملنا أيضاً على ضمّ القيم التربيعية - تفاعلات متغير ما مع نفسه - للسماح بعلاقات منحنية الأضلاع بين متنبئ X، والنتيجة Y. وإذا ما قررنا أيضاً السماح بتفاعلات من ثلاث اتّجاهات، (لنقل، بالعمر، والجنوسة، والدخل)، فسيصل عدد المتنبئات في النموذج إلى 320 متغير. ومع ذلك، ليس كُلّ تلك 320 متغير - على ما يبدو - مهمة إحصائياً، أو متنبئات ذات دلالة أو مهمة. إذن، بدأنا نشعر بجاذبية خوارزمية اختيار سمة ما، التي تستطيع أن تخبرنا بالمتغير المهم من بين هذه 320 متغير.

وتعمل إجراءات الانحدار التدريجي من خلال إحدى الطرق الثلاثة، بحيث تبدأ

الأولى - الانتقاء الأمامي (Forward Selection) - بنموذج يضمّ فقط متغير اعتراض (Intercept). وبعد ذلك تفحص كُلّ متغير مستقل على حدة، وتختار «الأفضل» (وسنعود إلى كيفية تحديد الأفضل بعد حين). وبعد دخول هذا المتغير في النموذج التنبؤي، يعيد البرنامج هذه العملية، مع اعتبار المتنبئات المرشحة المتبقية مراراً وتكراراً - بإضافة متنبئ متفوق في الوقت نفسه - إلى حين اختيارها النموذج «الأفضل» (ومرة أخرى سنحدد ذلك لاحقاً).

أما الطريقة الثانية للإزالة الراجعة (Backward Elimination)، فتبدأ بضمّ كُلّ المتغيرات المتاحة في انحدار أولي، وبعدها تختبر كُلّ واحد، للنظر في المتغير الذي يمكن أن يكون إقصاؤه من النموذج أمراً مفيداً. وتنتهي بنموذج أكثر انخفاضاً من حيث عدد المتغيرات «المعتمدة». **وأخيراً**، ثمة طريقة معروفة بالانحدار التدريجي الأمامي - الراجع (Forward–Backward Stepwise Regression) تجمع بين الانتقاء الأمامي والإزالة الراجعة، كما يبين الاسم ذلك. ومثلها في ذلك مثل الانتقاء الأمامي، فالطريقة تبدأ بنموذج صفري/ عدمي (Null Model)، وتدخل متغيرات بشكل تكراري عندما تلبي معياراً ما، ولكنها أيضاً تزيلها (في حالة ما) وعندما تنزل بعد ذلك تحت عتبة مناسبة.

ويتم انتقاء المتغيرات سواء على مستوى ضمها أو إقصائها من خلال إحدى الطريقتين: أما الطريقة الأولى، فتضم استخدام قيم- p بالنسبة إلى متغيرات المتنبئ الفردي. مثلاً، قد يعطي الباحث تعليماته للبرنامج بضمّ متغيرات فقط إذا كانت تتوافر على قيم-p، تصل إلى 0.05 أو أقل من ذلك، وتقصيها إذا ما تجاوزت قيمتها 0.10. وإن معيار إدراج المتغير هذا، الموجه بالكامل نحو المتنبئات الفردية، هو ما يستخدم حصرياً، في الخوارزميات التدريجية لبعض الحزم التجارية مثل حزم «الستاتا» (Stata).

وثمة مقاربة بديلة تتمثل في استخدام قياس ما، من قياسات تناسب النموذج الشاملة أو العامة، بحيث عادة ما يوجد قياس يَعْقُب نموذجاً ما، لإضافته مزيداً من التنبؤات. كما تضم قياسات التناسب، R^2 معدَّلة، ومعيار أي آي سي للمعلومة (AIC)، ومعيار بايز للمعلومة (BIC)، وقيمة «مالوز» C_p. ويتم ضمّ المتغيرات أو حذفها على أساس التحسن الذي يقوم به كُلّ متغير لفائدة نموذج الانحدار بشكل عام، كما تم تقييمه من قبل التغيير في الإحصائيات التناسبية المعينة من لدن الباحث.

يتم اختيار «النموذج النهائي» بطريقة تشبه طريقة اختيار المتغيرات الفردية. وإذ ما تم استخدام قيم-p في انتقاء المتغير، فستتوقف الخوارزمية في بناء نموذج انحدار تدريجي بمجرد وجود كُلّ المتغيرات التي تستجيب للمعايير المحددة للباحث في النموذج (مثلاً، كُلّ المتغيرات في النموذج لديها قيم -p 0.05، أو أقل، ولا يدخل أي متغير آخر في النموذج الذي قد يكون له هذا النوع من قيمة-p. ومن ناحية أخرى، إذا تم استخدام إحصاء تناسبي عام، فستختار الخوارزمية النموذج الذي يحسّن ذلك الإحصاء التناسبي - أي النموذج الذي يملك أعلى قيمة معدَّلة، أو معيار بايز للمعلومة الأقل انخفاضاً.

وكلا الطريقتان سريعتا التأثر بخطأ النوع 1، لأنه لا يتأثر - على الأرجح - بإيجاد خطأ واحد، المرتبط بالنتيجة بشكل كبير، في مجموعة متنبئات كبيرة، إلا بمحض الصدفة. وبينما قد يعد هذا واضحاً جداً في الحالة المتعلقة بالطرق القائمة على قيمة-p للانتقاء، ينطبق الأمر أيضاً على طرق الانتقاء التي تستخدم مقاييس عامة لتناسب النموذج. وبالنظر إلى وجود متنبئات كافية، سيكون لزاماً على المرء - بمحض الصدفة - رفع القدرة التنبؤية بشكل كافٍ لتجاوز عتبة الإدراج (الضمّ).

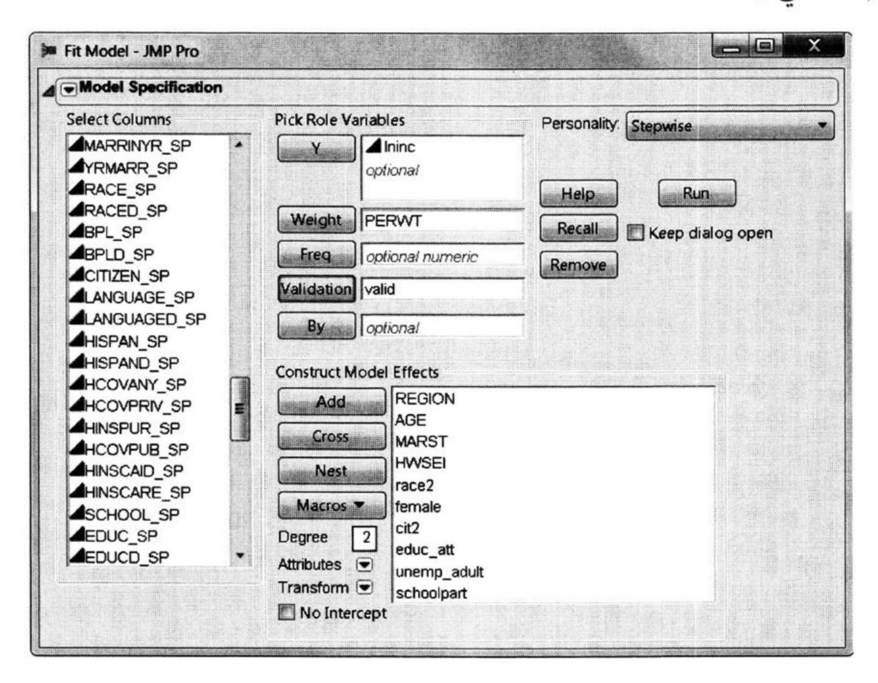
ويبدو أن أفضل الطرق لتجنب خطأ النوع 1، هي تلك الطرق التي تستخدم قياساً لتناسبية النموذج باعتباره «قاعدة توقف»، والتي تنتقي المتغيرات على أساس قيم-p، ولكن قيم-p التي تأخذ بعين الاعتبار قضية المقارنات المتعددة. وقد نطبق مثلاً قاعدة بون فيروني (Bonferroni Rule)، محددين قيم-p في α/p، حيث إن p هي مجموع عدد متنبئات مرشحة و α هي 0.05. ومع ذلك، إن قيم-p بون فيروني تعد صارمة جداً. وثمة مقاربة أخرى اقترحها (Foster and Stine 2004)، تفيد باختيار متغيرات في ترتيب تصاعدي بحسب إحصائيات اختبارها (t Statistics)، بدء بعتبة متحفظة، ورفع ذلك العتبة تدريجياً، بالتزامن مع العمل نحو متغيرات تنبؤية أقل.

ومثلها مثل أي مقاربة أخرى تعتمد على البيانات، سيكون من المرجح جداً أن يفرط الانحدار التدريجي في تناسبية النموذج قيد الدرس (ولو أن استخدام عتبة أكثر صرامة للإدراج، سيحل هذا إلى حدّ ما). ومن ثم، يعد التحقق من البيانات انطلاقاً

من مجموعة اختبار منفصلة للبيانات، أمراً مهماً. ولابد من نقل تناسبية الصلاحية المتبادلة وقت ما كان ذلك ممكناً.

مثال في «الغامب برو»

سنبين أهمية استخدام الانحدار التدريجي للخوارزمية التدريجية «للغامب» التي نرغب فيها بسبب الخيارات المتعددة لقواعد التوقف التي تمنحها، وبسبب السهولة التي يمكن أن تضاف معها التفاعلات والمتغيرات المتعددة الحدود (Polynomial) إلى النموذج. (ومع ذلك، إن الانحدار التدريجي متاح في حزمات إحصائية أخرى عديدة بما في ذلك الحزمة الإحصائية للعلوم الاجتماع (SPSS)، ونظام التحليل الإحصائي (SAS)).

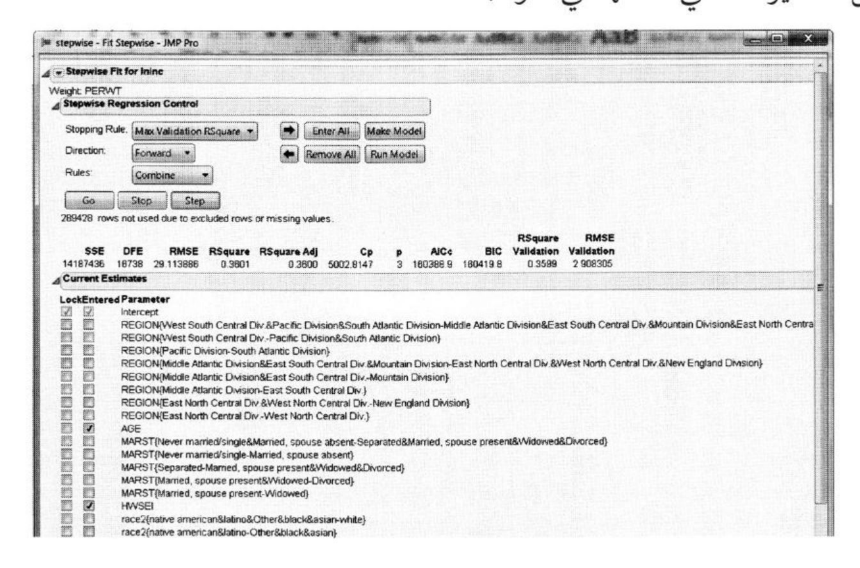


الشكل رقم 1.6: الانحدار التدريجي في «الغامب برو».

سنستعمل في هذا الشرح، بيانات صادرة عن مكتب تعداد مسح المجتمع الأميركي لعام 2010، التي سحبنا منها بشكل عشوائي مجموعة بيانات تهم 15000 حالة، وسنقوم بعملية تنبؤ الدخل الشخصي (المسجل) للبالغين في الولايات

المتحدة، وذلك باستخدام عدد صغير - إلى حدّ ما - لمتغيرات التأثير الأساسية: المنطقة، العمر، الحالة الاجتماعية، والاعتبار المهني، والعرق، والجنوسة، والمواطنة، والتحصيل العلمي، والحالة الوظيفية، والالتحاق بالتعليم.

ولتشغيل انحدار تدريجي، مستخدمين «الغامب» نتوجه إلى «حلّل» (Analyze)، ونختار نموذج التناسب (Fit Model)(الشكل رقم 1.6). وفي الزاوية العليا اليمنى لعلبة حوار نموذج التناسب (Fit Model Dialog Box)، ننقر القائمة الشخصية (Personality Menu)، ونختار «متدرج» (Stepwise)، كما نستطيع إضافة ترجيح احتمالية (متغير يدعى PERWT، المقدم من قبل المسح لتصحيح عدم الاستجابة). ونخبر البرنامج أيضاً باسم متغير صلاحية ما، الذي سميناه «صالح» وقمنا بإنتاجه سابقاً في «الغامب» ونقسم عشوائياً، مجموعة البيانات إلى مجموعة بيانات التدريب ومجموعة اختبار (في نسبة 2:1). وبعدها ننقر شغّل (Run)، وهذا يفتح منصة الإطلاق التدريجي (Stepwise Launch Platform) (الشكل رقم 2.6)، التي تعدد كُلّ المتغيرات التي نضمها في نموذجنا.



الشكل رقم 2.6: مُخرج من الانحدار التدريجي في «الغامب برو».

يجب علينا تفسير قيام «الغامب» بشيء ذكي ذي متغيرات فئوية، غير ثنائية التفرع

في انحدار تدريجي. وبدلًا من عرض المتغير الفئوي (Categorical Variable)، باعتباره مجموعة متغيرات وهمية لصفر أو واحد، مع حذف فئة خط أساسي واحد، يقوم «الغامب» بترميز الفئات تراتبياً. ويقسم الفئات أولاً إلى مجموعتين لهما وسائل متفاوتة بالنسبة إلى متغير الاستجابة، ثم يضيف متغيراً وهمياً لهذه المفارقة. وداخل هذين المجموعتين، تقسمها بعد ذلك مرة أخرى إلى مجموعتين أُخريين على النحو نفسه، وهكذا. على سبيل المثال، تأمل ما قام به «الغامب» مع متغير «التحصيل العلمي»؛ فهو يقسمه أولاً إلى مجموعتين: أقل من المستوى الثانوي، مقابل كُلّ الفئات الأخرى، ثم بعد ذلك يقسم هذه المجموعة الأخيرة إلى مدرسة ثانوية + كلية ما + غياب أي درجة علمية، مقابل درجة الزميلة + درجة الباكالوريوس + مستولى أعلى من درجة الباكالوريوس:

- أما المجموعة الأولى، فبإمكانها القيام بعملية التقسيم مرة واحدة،

- في حين إن المجموعة الثانية تنقسم مرة أخرى إلى درجة الزميلة + درجة الباكالوريوس مقابل مستوى أعلى من درجة الباكالوريوس.

وتتمحور الفكرة في كون أن هذه المجموعات تم تجميعها تراتبياً على مستوى النتيجة بحيث تبقى المجموعات ذات معدلات أكثر تماثلًا على مستوى متغير النتيجة، مجتمعة ضمن مجموعة واحدة. وهذا الاختلاف التراتبي لديه فائدة السماح للبرنامج باختيار نموذج انحدار أكثر تقتيراً من نموذج يضم كُلّ القيم المنفصلة لمتغير فئوي باعتبارها متغيرات وهمية إذا كان ذلك التقتير مفيد لتناسب النموذج. وكما هو مبين في لقطة الشاشة (الشكل رقم 2.6)، فلقد حددنا متغير صلاحية، وبالتالي سنستخدم الحدّ الأقصى لصلاحية R^2 كقاعدة توقف، وعتبات قيمة p- (تدخل النموذج وتغادره)، والحدّ الأدنى لمعيار (AIC) أكايكي للمعلومة والحد الأدنى لمعيار بايز (BIC) للمعلومة. ثم نحدد الاتّجاه الذي يجب على الخوارزمية التدريجية المضي قُدماً فيه، ونختار الانحدار التدريجي. أما الخيارات الأخرى، فتدعى الخيار الراجعة (Backwards)، والخيارات المختلطة (Mixed)، بحيث يكون هذا الأخير متاحاً، فقط عندما تستخدم قواعد التوقف لقيمة-p.

إن قائمة القواعد (Rules) ترتبط بمتغيراتنا الفئوية المنظمة تراتبياً، إذ إن في الإعداد الافتراضي (Default Setting)، وادمج (Combine)، والخيار «المقيد»، سيثير انتقاء تمايز بالغ الدقة، إدراجاً تلقائياً لمجمعات رفيعة المستوى. وإذا لم تكن ترغب في ذلك، غير الإعداد إلى «لا قواعد» (No Rules) (وهو الأمر الذي لا ننصح به البتة). وعلى نحو عرضي، تظهر النماذج ذات متغيرات التفاعل على نحو مماثل في خوارزمية تدريجية «للغامب»؛ أي إن انتقاء تفاعل ما، سيؤدي تلقائياً إلى إدراج متغيرات مكونة كلها، اللهم إلا إذا لم يتم انتقاء «لا قواعد» (No Rules).

يسمح لك «الغامب» بتشغيل البرنامج خطوة واحدة في كُلّ مرة لمعاينة تطور النموذج. أما الخطوة الأولى في نموذج التأثيرات الأساسية فقط، فتضم الاعتبار المهني، وتبلغ R^2 من 0.21. وتضم الخطوة الثانية العمر الذي يرفع R^2 إلى 0.36، متبوعاً بمؤشر بالنسبة إلى بالغين عاطلين في سنّ العمل (R^2=0.44). وعندما نسمح بتشغيل الانحدار التدريجي إلى النهاية، واختيار النموذج الأفضل لتنبؤ الدخل الخوارزمي، تستمر الخوارزمية في انتقاء 24 معلم، وتبلغ صلاحية R^2 من 495. كما هو مبين في الجدول رقم 1.6. ويمكن رؤية تطور تناسبية النموذج، من خلال اختيار - في أعلى الزاوية من ويندوز المُخرَج - قائمة «المثلث الأحمر»، تاريخ المعيار (Criterion History)، ثم تاريخ مربع R (R-Square). (ويحتوي العديد من ويندوز «الغامب»، قوائم تدعى المثلث الأحمر الذي يشير إلى الأسفل). وعلى الرغم من أن هذا الكتاب تم طبعه بالأسود والأبيض، سنظل مع ذلك نشير إلى هذه القوائم باعتبارها مثلثات حمراء). وكما هو مشار إليه في الشكل رقم 3.6، إن معظم التحسن الواقع في تناسبية النموذج، تم بلوغه في العشر خطوات الأولى، ولم يتم بلوغ تحسينات متواضعة جداً إلا بعد عشرين خطوة أو متنبئات. ومع ذلك، يستمر التناسب في التحسن تدريجياً.

بكل تأكيد، نستطيع القيام بأفضل من هذا من حيث التناسب، إذا ما أدرجنا متغيرات تفاعل، من خلال اختيار نموذج الإعادة (Relaunch Model)، الذي تعيدنا إلى علبة نموذج التناسب. ونضع الآن الدرجة (Degree) في 2 (بالنسبة إلى متغيرات تفاعل من اتّجاهين). وفي علبة اختر الأعمدة (Select Colums)، نبرز كُلّ متغيراتنا،

ثم نختار من قائمة ماكروس (Macros)، (Factorial to Degree). وتدخل هذه تلقائياً كُلّ التفاعلات الممكنة في اتّجاهين باعتبارها متغيرات مرشح. ومن أجل قياس جيد، ندرج أيضاً متغيرات تربيعية بالنسبة إلى العمر، والاعتبار الوظيفي. وفي هذه المرة، يأخذ البرنامج 54 خطوة لبناء النموذج الأمثل على مستوى تناسب الصلاحية. وأصبح للنموذج المنتقى حالياً R^2 من 0.6123 في مجموعة التدريب، و0.6064 في مجموعة الاختبار، وأدخلت 68 معلم في النموذج، بما في ذلك متغيرات التفاعل.

وتعد العديد من المتغيرات المنتقاة تفاعلات، تثير إدراج المتغيرات المكوّنة. ومن ثم، فإن كُلّ متغيرات التأثير الأساسي، استعملت - إلى حدّ ما - في النموذج، ولكن ليس كُلّ الفئات المنفصلة للمتغيرات الاعتبارية (Nominal) أو الفئوية تم استخدامها. على سبيل المثال، لم يتم إدراج إلا ثلاثة مناطق تباينات، ولم يتم تصنيف متغير التحصيل العلمي بالكامل.

وفي نموذج من هذا التعقيد حيث استخدام العديد من التفاعلات، إلى جانب تجمعات الفئوية، يصبح تفسير المَعْلمات أمراً صعباً. وإذا ما درسنا مَعْلم العمر، مثلًا، فسنجد أن النموذج قد اختار التأثير الأساسي للعمر، والمتغير التربيعي، وثمانٍ متغيرات تفاعل تتضمن العمر. وهنا تظهر مقايضة قوية بين الدقة التنبؤية وقابلية التأويل. وبهذه الدرجة من التعقيد، يكون من الصعب - وإن كان غير مستحيل - تأويل فقط ما سيكون عليه «تأثير تغيير وحدة واحدة في العمر على الدخل». ومع ذلك، إن إدراج متغيرات التفاعل هذه قد زاد من الدقة التنبؤية الخارجة عن البيانات.

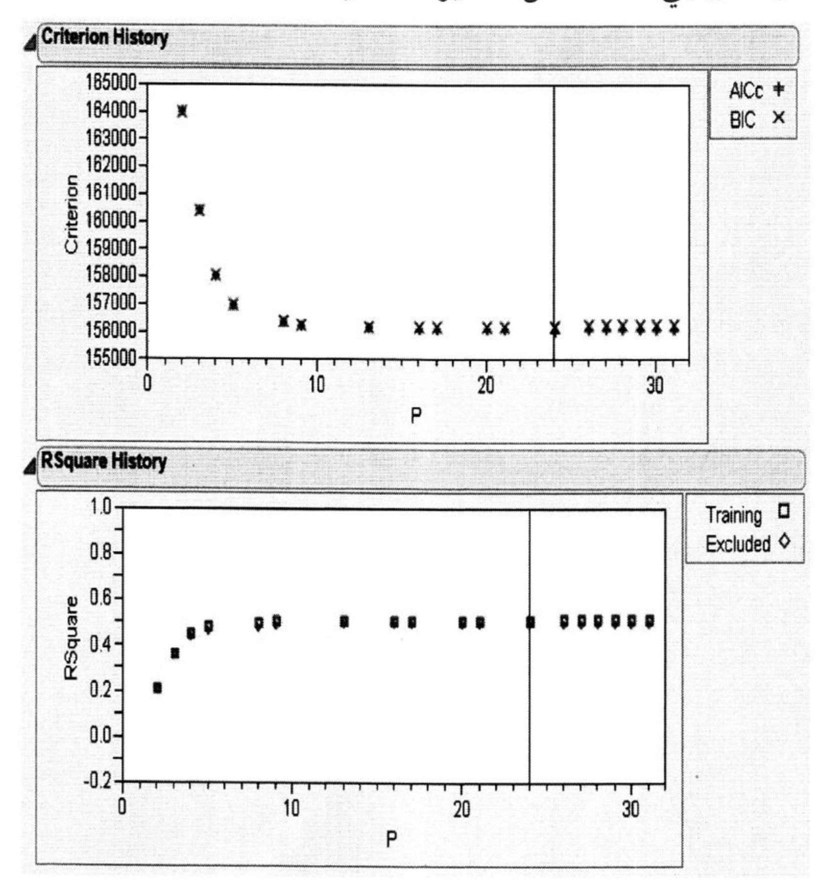
الجدول رقم 1.6: نموذج انحدار التأثيرات الأساسية.

	β	p
المنطقة: غرب جنوب الوسط، جنوب المحيط الأطلسي والهادي، مقابل أخرى	−0.030	.011
المنطقة: غرب جنوب الوسط، مقابل جنوب المحيط الأطلسي	−0.097	.001
المنطقة: جنوب المحيط الأطلسي، مقابل الهادي	0.075	.028

المنطقة: منتصف المحيط الأطلسي، جنوب شرق الوسط، الجبل، مقابل شمال شرق الوسط، شمال غرب الوسط، وبريطانيا الجديدة	-0.023	.049
المنطقة: شمال شرق الوسط، شمال غرب الوسط، مقابل بريطانيا الجديدة	0.038	.104
المنطقة شمال شرق الوسط، مقابل شمال غرب الوسط،	0.085	.071
العمر	0.040	<.001
الحالة الاجتماعية: لم يسبق له الزواج قط، متزوج، الزوج غائب، مقابل منفصل، متزوج، الزوج حاضر، أرملة وطالق	0.062	<.001
الحالة الاجتماعية: متزوج، الزوج حاضر وأرملة، مقابل طالق	-0.133	<.001
الحالة الاجتماعية: متزوج، الزوج حاضر، مقابل أرملة	189.-	<.001
الاعتبار المهني	0.061	<.001
العرق: آخر، مقابل أبيض العرق: أميركي أصلي ولاتيني، مقابل أسود، آسيوي، وآخر	0.046	<.001
العرق: آخر وأسود، مقابل آسيوي	0.193	<.001
أنثى	0.261	<.001
غير مواطن	-0.281	.007
التعليم: أقل من درجة التعليم الثانوي، مقابل الباقي	-0.639	<.001
التعليم: المدرسة الثانوية، وكلية ما، مقابل AA وBA وأعلى من ذلك	0.141	<.001
التربية: التعليم الثانوي مقابل بعض الكليات	-0.146	<.001
التعليم: AA وBA وأعلى من ذلك	0.193	<.001
عاطل	-1.435	<.001
في المدسة	-1.039	<.001
ثابت	3.204	<.001
R^2	0.508	
صلاحية R^2	0.495	

خلاصة

يمكن استعمال الانحدار التدريجي في انتقاء - من أصل مجموعة كبيرة من المتغيرات المستقلة - تلك المتغيرات المستقلة الأكثر تنبؤاً. وعموماً فهو يستعمل في سياقات حيث لدى باحث ما سمات في مجموعة البيانات. ويمكن أيضاً استعمال التقنية لتحديد تلك المتغيرات التفاعلية بين المتنبئات التي تحسن القوة التنبؤية لنموذج ما. وعادة، ثمة العديد من متغيرات التفاعل الممكنة، ويمكن استعمال الانحدار التدريجي للكشف عن المتغيرات المفيدة.



الشكل رقم 3.6: تطور تناسبية النموذج التدريجي في «الغامب برو».

ومن الأفضل استخدام الانحدار التدريجي بنوع من الصلاحية المتبادلة، لأن

هذه التقنية ستفرط في تناسبية البيانات. ومن ثم، فإن R^2، أو الإحصائية التناسبية لبيانات التدريب ستكون عالية بشكل مصطنع. ومع ذلك إذا اهتم الشخص بالقوة التنبؤية لنموذج ما من أجل عينة الاختبار، فإن التفريط في التناسب لن يصبح مشكلاً. وإن R^2 بالنسبة إلى مجموعة بيانات الاختبار هو مقياس صالح للقوة التنبؤية لنموذج الانحدار.

اللاسو

كما تمت الإشارة إلى ذلك، ينتقص بعض المحللين من شأن الانحدار التدريجي، مثلما ينتقصون من شأن العديد من مقاربات التنقيب في البيانات، والذين يرون الطريقة باعتبارها «تجريفاً للبيانات» (Data Fishing or Data Dredging) غير نظرية. ويرى هؤلاء الباحثون أن تحرك النمذجة الإحصائية نظرية حول العمليات السببية والمتغيرات التي تمثلها. ولكن الانحدار التدريجي تعرض أيضاً للنقد من داخل الحقل المعرفي نفسه الذي يبحث في التنقيب في البيانات. ويشير هؤلاء النقاد إلى أن طبيعة عملية انتقاء متغير - أي ضمّ المتغيرات أو تركها - يجعل الانحدار التدريجي غير مستقر، ومن ثم غير موثوق به إلى حدّ ما. وإن التغييرات الصغيرة في البيانات، مثل معاينات (Samplings) عشوائية مختلفة مأخوذة من مجموعة أكبر من الحالات، يمكن أن يقود إلى اختيار مجموعات فرعية من المتغيرات من لدن خوارزمية تدريجية. والطريقة التي عوض أن تحتفظ بالمتغيرات بالجملة أو تتخلص منها، تقوم بانتقاء أكثر تدرجاً واستمرارية، تبدو مفضلة.

إن «اللاسو» (الذي يشير إلى الانكماش المطلق للغاية، وإلى مشغل الانتقاء)، يمثل هذه الطريقة بالذات. وتفرض جزاء على معاملات انحدار النموذج على نحو تنكمش فيه تلك المتغيرات الأقل تنبؤاً، نحو الصفر. وهذا يجعل «اللاسو» مماثلاً في الشكل للإزالة الراجعة (Backwards) للانحدار التدريجي. وخلافاً للإزالة الراجعة، كما إن تبسيط النموذج في «اللاسو» لا يحدث عبر تأسيس عتبة عشوائية. وإن الطبيعة التدريجية لعملية انكماش «اللاسو»، يعني أن إدراج متغير ما أو إقصاءه، لا يؤثر بشكل مباشر وعميق في معاملات تلك التي بقيت. ومن ثم، فإن «اللاسو» أكثر استقراراً، وينتج تحيزاً أقل من الانحدار التدريجي، ولكن مع ذلك تنتج تبسيط نموذج مماثل.

ورياضياً، يرتبط جزاء «اللاسو» بمجموع القيم المطلقة لمعاملات الانحدار

(وهذه هي مسافة «مانهاتن» أو تجمع المدينة للقوة الموجهة للمعاملات، التي تدعى أيضاً معيار L_1)، عادة بعدما يتم تقعيد كُلّ المتنبئات باعتبارها الفرق المعياري عن متوسط القيم z- (Z-Scores). وفي صيغتها الأولى (Tibshirani, 1996)، تم تقييد مجموع هذه القيم المطلقة لتصبح أقل من معلم تضبيط، t. وإذا حُدّدت t أكبر أو تساوي المجموع المرصود للقيم المطلقة للمعاملات من نموذج المربعات الصغرى العادية (OLS) للخط الأساسي (Baseline)، فلن يحدث أي انكماش وستساوي تناسبية «اللاسو». وإن لدى عملية تقليص معلم التضبيط هذا إلى أدنى من ذلك المجموع، تأثير تقييد هذه المعاملات. وتستخدم تباينات أخرى - مثل الطريقة التي نستخدم - تحولاً لهذا المعلم الذي يزيد من القيود أكثر في أعلى قيم.

ولبيان ذلك، نبدأ بعرض نموذج مربعات صغرى عادية بشكل كامل، متنبئين بنسبة الأصوات لدى أوباما في العام 2012 في محافظات الولايات المتحدة. وهذا الانحدار «العادي» سيستخدم كمؤشر مرجعي (Benchmark) نقارن من خلاله «اللاسو». واخترنا هنا مجموعة كبيرة إلى حدّ ما، من المتغيرات المستقلة - 22 في المجموع - التي تصف أبعاد ديموغرافية متعددة لهذه المحافظات (الكثافة السكانية، والمزيج العرقي، وبنية العمر، والخصائص الاقتصادية، وغيرها). وقد بيّن الجدول رقم 2.6 نتائج هذا الانحدار. وللتيقن من أهمية هذا النموذج في حدّ ذاته، وأنه يفسر نسبة جيدة من التباين في التصويت: 58٪. ولكن، وبشكل واضح، لدينا بعض المتنبئات المترابطة، وقد نهتم بنموذج أكثر تجرداً وتقتيراً. وتعد هذه مناسبة ممتازة لاستخدام «اللاسو». وثمة برنامج «الستاتا»، المقدم من لدن المستخدم الذي ينفذ «اللاسو» («اللارس» (Lars) يحدد وظيفة «اللاسو»)، ولكن يظهر أنه في بداية مراحله من التطور. وتوجد القدرة من أجل «اللاسو» في «الغامب برو» 12، وفي إحصائيات الحزمة الإحصائية للعلوم الاجتماعية (SPSS)، مادام يشتري الشخص حزمة فئات الحزمة الإحصائية للعلوم الاجتماعية. ولدى R - على الأقل - روتينان (Two Routines) ينجزان «اللاسو»، ويسميان «بينلايزد» (Penalized)، و«لارس»، وهما متاحان عبر الرابط: cran.rproject.org. وسنفترض ألفة خط أساسي مع R، مع التركيز هنا على حزمة «بينلايزد» (Goeman, 2010; Goeman, Meijier and Chaturverdi 2012).

ونبدأ بدعوة بسيطة لدالة R، بما في ذلك الإعدادات الافتراضية في الغالب، للبرنامج:

```
lassol < -penalized (obama~Inpopdens+agelt18+age1834+age65over+
  imdens+perwhite+perasian+perblack+perlatin+edhigher+edhs+edl
  hs+unempmale+unempfem+perpov_q+divorce2per+samesexper+e
  vprot10+hhsize+occprofman+medinc+hsdrop1619, lambda1 = 500,
  standardize = TRUE)
```

الجدول رقم 6-2: نتائج من انحدار مربعات صغرى عادية تتنبأ بحصة أوباما من الأصوات ضمن بيانات على مستوى المحافظة.

المتغير	المعامل (SE)	المعامل المقعّد
الكثافة السكانية (log)	2.398 (0.157)***	0.278
٪ أقل من 18 عاماً	-0.775 (0.101)***	-0.177
٪ بين 18-34 عاماً	-0.534 (0.0710)***	-0.177
٪ 65 وأكثر	-0.636 (0.100)***	0.179
٪ البيض من غير الإسبان	-0.476 (0.0311)***	-0.625
٪ آسيوي	0.165 (0.116)	0.0260
السود من غير الإسبان	0.0147 (0.0309)	0.0145
٪ لاتيني/ a	-0.0748 (0.0360)**	-0.0651
٪ خريج كلية	0.563 (0.0613)***	0.329
٪ خريج ثانوية فقط	0.314 (0.0477)***	0.147
٪ مستوى أقل من التعليم الثانوي	-0.214 (0.0514)***	-0.106
معدل البطالة لدى الرجال	0.920 (0.0729)***	0.232
معدل البطالة لدى النساء	0.0808 (0.0795)-	-0.0184
معدل الفقر	0.249 (0.0642)***	0.101
٪ مولود بالخارج	-0.174 (0.0643)***	-0.0637

٪ طالق	0.104 (0.0424)-**	-0.0438
٪ أسرة مكونة من الجنس نفسه	1.718 (0.582)***	0.0374
٪ بروتستانتي أنجليكاني	-0.277 (0.0133)***	-0.304
معدل حجم الأسرة	-8.242 (1.341)***	-0.135
٪ مهني/ إداري	-0.369 (0.0487)***	-0.163
متوسط الدخل	-7.34e-05 (4.16e-05)*	-0.0567
معدل الهدر المدرسي	-0.0173 (0.0337)	-0.00655
ثابت	130.0 (7.903)***	
ترصدات	3,114	
R^2	0.586	

I J OE: الأخطاء المعيارية في القوسين.
***P < .001, **p < .01, *p < .05

وهذا يشغل النموذج. ويحدد خيار «لامدا» 1، الجزاء المرتبط بمجموع القيم المطلقة للمعاملات. وستنتج قيم أكبر انكماشاً أكثر نحو صفر معاملات الانحدار. ومن الممكن أيضاً استعمال خيار منفصل يدعى «لامدا» 2، المرتبط بجذر المربع لمجموع مربعات معاملات الانحدار (مسافتها الإقليدية أو معيار L_2). وسيؤدي إدراج لامدا 2 عوض لامدا 1، «بينلايزد» إلى إنجاز انحدار الحيْد (Ridge Regression). ومن الممكن في «بينلايزد» تحديد كُلّ من لامدا 1 ولامدا 2 لجزاء النموذج على نحو أكثر تعقيداً. وقد يقصي المرء أيضاً بعض المتغيرات المشاركة من الجزاء، وقد يجزي المعاملات المتنوعة بشكل مختلف. ولكن سنركز هنا على حالة مباشرة من اللاسو. وقد قعّدنا أيضاً متغيراتنا باعتبارها فرقاً معياري عن متوسط قيم-z، مقدَّماً (بحيث يكون standardize = TRUE).

ولرؤية معاملات الانحدار، ندخل

Coefficients (lasso1, «all»)

يظهر جدول رقم 3.6 نتائجنا. ونحدد ابتداء مَعْلم جزاء منخفض هنا (قيمة 2 بالنسبة إلى لامدا 1): انظر العمود المسمى 2. وبالتالي، فإن لدى كُلّ المتغيرات معاملات لا صفرية/ عدمية، وهي متطابقة تقريباً مع تقديرات المربعات الصغرى العادية. وفي الحقيقة، علينا جعل الجزاء أكبر للغاية لرؤية تغيير نموذج قوي. ولا شيء يُسقط بتاتاً إلى أن يبلغ الجزاء 500. وفي العمود الذي يضمّ 500، تنخفض بعض المعاملات بالنسبة إلى المتنبئات إلى الصفر. حتى بعد مضاعفة الجزاء مجدداً (إلى 1,000)، نحتفظ بـ 15 متغير مشارك. ويحدث هذا من دون شكّ، لأن العديد من متغيراتنا المشاركة، تساهم – في الحقيقة – في تفسير التباين في النتيجة، وبسبب حجم عينتنا الكبيرة (N = 3,114) نسبياً.

وبمجرد أن تبدأ المتغيرات في الانكماش إلى الصفر، تحدث بعض الأشياء المهمة؛ فبينما معظم المعاملات تنكمش بشكل مفردة النغمة بارتفاع الجزاء، يرتفع المعامل في نسبة السود إلى أن يساوي لامدا 1 = 1,000، وبعدها تنخفض قليلاً. وأما المُعامل الصغير في حصة السكان السود في النموذج الأول، فقد كان مفاجئاً. وهذا يقترح أن في نموذج متعدد التغيرات، تكون تأثيرات هذا المتغير مقنعة بالمتغيرات المشاركة المتصلة، ولكن يمثل هذا متنبئاً مهماً في حدّ ذاته ولذاته. وأما المتغير المتعلق بالناس المطلقين باعتبارهم نسبة تمثل بالغين لم يتزوجوا قط، فينخفض إلى الصفر في لامدا 1= 500، ويعاود الظهور في 1000، ثم ينكمش إلى الصفر. وفي العمود الأخير من الطباعة، ذي جزاء يبلغ 5,000، لدينا مجموعة أصغر جداً من متغيرات مشاركة للفحص، بحيث يفسر كُلّ واحد منها قدراً متواضعاً من التباين في البيانات.

يستطيع «بينلايزد» إنتاج رسم بياني، مبينا كيفية انكماش معاملات الانحدار بالتزامن مع ارتفاع الجزاء. ولرؤية هذا الرسم البياني، نخبر **أولاً** البرنامج بغرض حساب المعاملات في الوقت الذي ترفع فيه العقاب على فترات منتظمة (الخطوات = 100). وبينما يكون بالإمكان جداً رسم بيان، باستخدام عدد أكبر من المتغيرات كما في النموذج أعلاه، فإن الرسم البياني المحصل عليه، سيكون لمسة مكتظة. ولأجل عرض واضح، نقدر نموذجاً أكثر بساطة:

```
lasso1<-penalized (Obama , ~ lnpopdens+imdens+perplack+perwhite+
  edhigher+evprot10, lambda1 = 2, steps = 100, trace = FALSE,
  standardize = TRUE)
```

الجدول رقم 3.6: معاملات الانحدار من «اللاسو» التي تتنبأ بـ حصة أوباما من الأصوات بعقوبات متفاوتة.

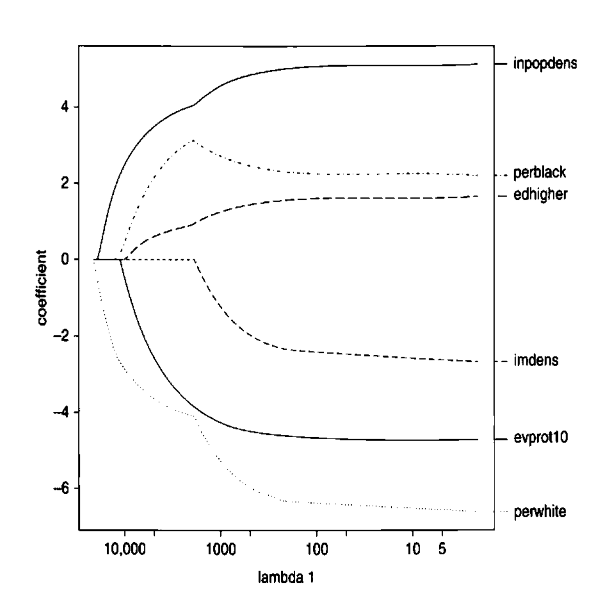
	قيمة جزاء لامدا 1				
	2	**100**	**500**	**1000**	**5000**
السكان/ sq.mile (log)	2.396	2.313	2.139	2.063	1.631
٪ العمر<18	-0.7730	-0.689	-0.474	-0.375	-0.131
٪ البيض غير الإسبان	-0.4756	-0.452	-0.372	-0.300	-0.172
٪ السود	0.01499	0.027	0.079	0.125	0.116
٪ الباكالوريوس أو درجة أعلى	0.5619	0.527	0.402	0.246	0.166
٪ نسبة البطالة بين الرجال	0.9195	0.916	0.973	1.025	0.699
٪ بروتستانتي أنجليكاني	-0.2767	-0.274	-0.263	-0.249	-0.163
٪ آسيوي	0.1646	0.150	0.101	0.0301	0.000
٪ دبلوم المدرسة الثانوية	0.3139	0.301	0.233	0.126	0.000
٪ أقل من دبلوم المدرسة الثانوية	-0.2140	-0.200	-0.171	-0.158	0.000
٪ معدل الفقر	0.2489	0.238	0.153	0.070	0.000
٪ أسرة مكونة من نفس الجنس	1.716	1.644	1.504	1.355	0.000
معدل حجم الأسرة	-8.242	-8.263	-7.425	-5.926	0.000

٪ المهني/ الإداري	-0.3682	-0.333	-0.201	-0.053	0.000
معدل الطلاق	-0.1036	-0.062	0.000	0.012	0.000
٪ العمر 18-34	-0.5319	-0.439	-0.161	0.000	0.000
٪ العمر 65+	-0.6331	-0.509	-0.177	0.000	0.000
٪ مولود بالخارج	-0.1735	-0.159	-0.087	0.000	0.000
٪ لاتيني	-0.07460	-0.063	0.026	0.000	0.000
معدل البطالة بين النساء	-0.07988	-0.036	0.000	0.000	0.000
معدل الدخل (s1000$)	-0.073	-0.030	0.000	0.000	0.000
معدل عدم إتمام التعليم الثانوي	-0.01715	-0.010	0.000	0.000	0.000

– ونقول له أن يقوم بإنتاج رسم بياني من هذه العلاقة. إنه يساعد على تقعيد المعاملات هنا لكي لا تؤدي القياسات المختلفة إلى أن يطغى بعضها على الآخر.

```
Plotpath (lasso1, log = «x», standardize = TRUE)
```

إن الرسم البياني المحصل عليه (الشكل رقم 4.6)، إضافة إلى التقعيد، تسمح لنا بتصور المتغيرات التي تبقى مهمة بالنسبة إلى النموذج. ونستطيع رؤية أن المتغيرات التي تبقى في النموذج مدة طويلة، هي تلك التي كانت – البداية – مترابطة بشكل كبير للغاية (إيجابياً أو سلبياً)، مع حصة أوباما من الأصوات: نسبة سكان المحافظة التي تمثل البيض غير الإسبان، والكثافة السكانية، ومعدل البطالة بين الرجال، ونسبة السكان الذين يمثلون البروتستانتيين الأنجلكان، ونسبة السكان البالغين الحاصلين على درجة الباكالوريوس، أو درجة أعلى. وإن لدى نموذج انحدار، يحتوي فقط على هذه المتغيرات الخمسة، R^2 معدلة من 0.49، مقارنة بـ .0.58 في النموذج بأكمله. كما يمكن أن تخبرنا هذه المتغيرات الخمسة، بعض الشيء، عن أنماط انتخابية كلية.



الشكل رقم 4.6: انكماش المعاملات

في إعدادات مختلفة من معلم الجزاء في «اللاسو» (من R).

ومع ذلك، إن إسقاط المتغيرات من النموذج - على الرغم من جعل النموذج أكثر قابلية للتأويل بشكل كبير - يقلص لا محالة من القيمة التنبؤية الإجمالية. وعموماً، نريد موازنة التقتير بدقة تنبؤية.

وتسمح لنا الحزمة المعاقبة أيضاً باستخدام الصلاحية المتبادلة لطية k- لتحديد مدى تناسبية نموذج ما. ويمكننا إنجاز الصلاحية المتبادلة لنموذج ما بالشفرة التالية:

```
cross<-cvl (obama, ~lnpopdens+agelt18+age1834+age65over+imdens+ perwhite+perasian+perblack+perlatin+edhigher+edhs+edlhs+edlhs+unempmale+unempfem+perpov_q+divorce2per+samesexper+evprot10+hhsize+occprofman+medinc+hsdrop1619, lambda1 =500, fold = 10, standardize = TRUE)
```

وينتج هنا شيء يدعى كروس (Cross)، الذي نخزن فيه نتائج الصلاحية المتبادلة،

انطلاقاً من نموذج يتنبأ بحصة أوباما من الأصوات، مستخدمين 22 متغيراً مستقلاً. وبعد النموذج، علينا تحديد معلم الجزاء (لامدا 1 = 500)، وعدد الطيات المستخدمة في الصلاحية المتبادلة (طية = 10). وبعد تشغيل النموذج، نقوم بدعوة عناصر الشيء. أما العنصر الأول – cvl$ – فيعيد الاحتمالية الخوارزمية للنموذج في بيانات الصلاحية المتبادلة. وبالعنصر fullfit$، يمكن دعوة التناسبية في البيانات بأكملها.

```
cross$cvl
cross$fullfit
```

الجدول رقم 4.6: إحصائيات التناسب

في قيم مختلفة لمعلم جزائي في اللاسو (LASSO).

	قيمة جزاء لامدا 1				
	2	**100**	**500**	**1,000**	**5,000**
احتمالية صلاحية التبادل	-11,483.17	-11,496.82	-11,516.6	-11,568.63	-11,888.41
احتمالية خواريزمية لبيانات كاملة	-11,443.92	-11,445.92	-11,475.19	-11,533.25	-11,827.07
معاملات لا صفرية	23	23	19	16	8

ونستطيع إعادة هذا، عدة مرات في إعدادات مختلفة من اللامدا 1، وفحص التناسبية النسبية للنموذج ذي معلمات جزاء مختلفة. وفي جدول رقم 4.6، نبين احتمالات خوارزمية للصلاحية المتبادلة بالنسبة للنموذج أعلاه، مع قيم اللامدا 1 التي استخدمناها أعلاه لبيان انكماش النموذج. ونرى أن الإحصائيات الدنيا للاحتمالية الخوارزمية المدرجة توجد في اللامدا 1 = 2. ويبدو أن هذا يقترح أن انكماشاً أقل – وليس أكثر – ينتج تناسبية أفضل في هذا النموذج. ولكن هذه النتائج القليلة، لا تسمح لنا باستنتاج مفاده أن 2 هو أفضل قيمة بالنسبة للامدا لتعظيم دقة خارجة عن العينة. وللقيام بذلك، علينا القيام بذلك كما تفعل خوارزمية ذات احتمالية قصوى، وذلك بتجربة قيم مختلفة، والتحرك أقرب فأقرب من الاحتمالية الخوارزمية الدنيا للصلاحية المتبادلة.

ويبدو أن هذا عملاً كثيراً، يجب أتمتته، ومن حسن الحظ، أن تم ذلك. وسيسمح لنا «بينلايزد» بإيجاد أفضل قيمة للامدا 1، بالدالة «optL1». كما تسمح لنا هذه الدالة بتحديد الحدّ الأدنى والأقصى لقيم لامدا، وستجد القيمة التي ستخفض الاحتمالية الخوارزمية للصلاحية المتبادلة إلى الحدّ الأدنى. ونحدد الحدّ الأدنى في 0، ونسمح للامدا 1 بأن يصل مداه إلى 1,000:

```
bestfit<-optL1 (obama, ~lnpopdens+agelt18+age1834+age65over+imdens+perwhite+perasian+perblack+perlatin+edhigher+edhs+edlhs+unempmale+Unempfem+perpov_q+divorce2per+samesexper+evprot10+hhsize+occprofman+medinc+hsdrop1619, minlambdal = 2, maxlambdal = 1000, fold = 10, standardize = TRUE)

bestfit$lambda
```

وهنا، بعد إنجاز 21 تكراراً، استقر البرنامج عند قيمة مثلى للامدا 1 هي: 27.14285. وكما تم ذكره آنفاً، في هذه القيمة المنخفضة، سيؤثر معلم الجزاء في معاملات انحدارنا تأثيراً ثانوياً، على الرغم من رفعها من الدقة التنبؤية قليلاً.

خلاصة

يعد «اللاسو» أداة انتقاء متغير قوي، يستخدم في إيجاد مجموعة فرعية لمتنبئات متاحة، لديها - مجتمعة - قوة تنبؤية. وإن المتخصصين في التنقيب في البيانات يستخدمونها لتحسين كُلّ من بساطة (تقتير) النموذج، والقوة التنبؤية. ولسوء الحظ، إن «اللاسو» ليس متاحاً بعد في بعض حزمات التنقيب في البيانات السهلة الاستخدام، ونتيجة لذلك، بينا التقنية باستخدام لغة R الحرة.

وفي المثال، استخدمنا «اللاسو» للتنبؤ بنسبة التصويت في محافظات الولايات المتحدة التي آلت إلى الرئيس أوباما في انتخابات 2012. وقد حدد البرنامج كثافة سكانية أعلى، ونسبة أقل من السكان دون السنّ 18، ونسبة أقل من البيض، ونسبة أعلى من السود، ونسبة تمثل درجات الكلية، ومعدل بطالة أعلى بين الرجال باعتبارها متنبئات جوهرية لحصة أوباما من الأصوات على مستوى المحافظة.

انحدار معامل تضخم التباين

يعد انحدار معامل تضخم التباين أداة أخرى من أدوات التنقيب في البيانات التي تم تطويرها حديثاً، من أجل تبسيط نموذج ما، من خلال انتقاء متغيرات (انتقاء سمة). وطور انحدار معامل تضخم التباين عام 2011 من قبل لين (Lin) وفوستر (Foster)، وأنغر (Ungar)، لاستخدامه تحديداً في مجموعات بيانات ضخمة جداً، خاصة تلك الكبيرة جداً (أعداداً كبيرة من المتغيرات). وقد تم تطويرها باعتبارها بديلاً عن الانحدار التدريجي وانحدار المجموعة الفرعية الأفضل، اللذين يعتبران مكثفين حاسوبياً، ومن ثم، يميلان إلى الاشتغال بشكل بطيء جداً. وأما طرق أخرى من طرق انتقاء السمة، مثل طالب الطريق المعمم (GPS)، فيشتغل على نحو أسرع بكثير، ولكنه يؤدي ثمناً في دقة تنبؤية متقلصة (Lin, Foster, and Ungar 2011). لقد صُمم انحدار معامل تضخم التباين، بغية تسريع الانحدار التدريجي دون مقايضة كبيرة على مستوى الدقة.

ويعد انحدار معامل تضخم التباين، خوارزمية متعددة المراحل، تمزج معاً، تقنيات كانت موجودة (مثل انحدار تدريجي أمامي، وقواعد استثمار ألفا)، وتضيف عنصرها الفريد. إنه اختلاف حول انحدار تدريجي أمامي (لأنه مع عدد كبير من سمات المرشح، تصبح الإزالة الراجعة غير كافية للغاية). وسنمر سريعاً على ما تقوم به الخوارزمية بالضبط، ونبين ما تستطيع تنفيذه باستخدام R.

إن الفائدة الرئيسة لانحدار معامل تضخم التباين، يتمثل في تقليص مقدار الحوسبة التي تحدث عند تشغيل انحدار تدريجي. ولكن، لماذا يتطلب الانحدار التدريجي حساباً رياضياً كثيراً؟ ثمة سببان اثنان وراء ذلك:

أولاً: لأن في كُلّ تكرار أو خطوة في بناء النموذج، تأخذ بعين الاعتبار كُلّ متغير مرشح من أجل إدراجه. وهذا يعني أن في كُلّ خطوة، يأخذ التدريج بعين الاعتبار عدة متغيرات، وبما أنها تنجز عدداً كبيراً من الخطوات، فهي تقوم بهذا عدة مرات. وفي المقابل، يدير انحدار معامل تضخم التباين متغيرات المرشح مرة واحدة فقط.

ثانياً: في كُلّ خطوة، يولّد التدريج تقديرات بالنسبة إلى كُلّ معلم. وهي لا تشغل

انحداراً واحداً فقط، وإنما العديد من الانحدارات بقدر وجود متغيرات المرشح في كُلّ مرحلة. ويلتف انحدار معامل تضخم التباين على هذا المشكل الثاني من خلال استعمال ما يسمى بالانحدار الأمامي على مراحل (Stagewise).

ويبدأ انحدار معامل تضخم التباين بنموذج صفري - أي بنموذج يضم فقط اعتراض (Intercept) واحد - ويحسب بقايا من هذا النموذج. ثم ينتقي المتغير الأول في قائمة المتنبئات المحددة سلفاً (التي يعد ترتيبها - إلى حدّ ما - أمراً مهما هنا)، وتعمل على تراجع هذه البقايا في هذا المتغير. وإذا ما استجاب المتغير إلى بعض المعايير التي تسمح بإدراجه، فسيدخل في النموذج، وتُحسب بقايا جديدة؛ وإلا، فإن الخوارزمية تنتقل إلى المتغير الموالي.

وبالتالي، إن في كُلّ خطوة على حدة، لا تتراجع إلا البقايا من المرحلة السابقة، في متغير المرشح الجديد. كما يسمح هذا الانحدار الأمامي على مراحل بتقليص كبير في الحوسبة. وعوض تشغيل انحدار كامل، وحساب كُلّ التقديرات المعلمية، يقوم انحدار معامل تضخم التباين - في الجوهر - بحساب فقط سلسلة من الارتباطات ذات المتغيرين (Bivariate).

ومع ذلك، هناك صعوبة بخصوص هذا الإجراء، ذلك بأن نسب-t (t-Ratios) المقيِّمة لهذه الارتباطات ذات المتغيرين، قد تتحيز ضدّ المتغيرات التي لها خطية متعددة (Multicollinearity) كبيرة مع المتغيرات التي سبق انتقاؤها للنموذج. ونتيجة لذلك، تتحيز خوارزمية انحدار تدريجي أمامي «ساذج» في اتّجاه متغيرات منتقاة غير مترابطة مع المتغيرات الموجودة سلفاً في النموذج عوض انتقاء ما قد يكون متغيراً تنبؤياً. من أجل هذا، لا بُدَّ من إحداث تصحيح ما، لإزالة هذا التحيز.

إن التصحيح هو ما يمنح انحدار معامل تضخم التباين اسمه؛ فانحدار معامل تضخم التباين، يعدل نسب-t لتفسير الخطية المتعددة من خلال استعمال «معامل تضخم التباين» لكُلّ متغير بما أنه يؤخذ بعين الاعتبار في هذا الإطار. ولكن بما أن انحدار معامل تضخم التباين، يتم حسابه بواسطة العمل على تراجع متغير جديد في كُلّ المتغيرات الموجودة سلفاً في النموذج، فإن ذلك يمثل لغزاً. وإن إنجاز هذه الانحدارات، سيزيل المدخرات في الحاسوب التي كانت الغاية وراء استعمال انحدار أمامي على مراحل في المقام الأول. ويتجلى الحلّ في تجنب حساب

معاملات تضخم التباين باستعمال مجموعة البيانات كاملة. وفي المقابل، تتم معاينة مجموعة فرعية صغيرة من الحالات بشكل عشوائي، وتقدر معاملات تضخم التباين من خلال هذا القدر الأصغر من البيانات.

وأخيراً، يسمح بتضخم احتمال خطأ النوع 1 بخصوص اختبارات فرضية متعددة. وبتعبير بسيط، عند إجراء اختبار لدلالة إحصائية، نسمح بإمكانية رفض خطأ، فرضية صفرية (العدم) حقيقية (خطأ النوع 1)، ثم نختار احتمال وقوع ذلك بتحديد ألفا. ولكن، كلما أجرينا اختبارات فرضية متعددة، تزداد احتمالية حدث نادر، أو ببساطة يعد تطبيق مستوى ألفا نفسها على كُلّ اختبار، غير مناسب.

وتقسم تصحيحات بونفيروني لاختبارات الفرضية المتعددة ببساطة، ألفا (عادة $p<0.05$) على n - عدد الاختبارات التي يجب إنجازها (أي عدد متغيرات المرشح). ولكن يطرح هذا التصحيح إشكالاً إذا كان عدد الاختبارات ضخمة، وترفع من احتمال خطأ من النوع 2 من خلال تحديد ألفا الفعالة من أجل الدخول، في مستوى منخفض جداً للغاية.

وفي واقع الأمر، إن انحدار معامل تضخم التباين، يستخدم إجراء، يدعى قاعدة استثمار ألفا (Alpha-Investing Rule) الذي يقيم حلاً وسطاً بين اختبار فرضية متعددة غير مفيدة (الذي ينتج العديد من الأخطاء من النوع 1)، وتطبيق قاعدة بونفيروني (التي تميل إلى إزالة متنبئات محتملة مهمة؛ انظر Foster and Stine 2008). وتتلخص الفكرة في كون أننا نبدأ «بثروة» معينة، أو بترخيص لخطأ النوع 1 (لنقل 0.05 أو 0.10). ثم نقوم بعد ذلك بإنجاز اختبار فرضية ما. وإذا ما تم رفض الفرضية الصفرية أو العدمية، فسنقوم بالزيادة في ثروتنا، وإذا أخفقنا في رفضها، فسننقص منها. وبالنتيجة، تُستنزف الثروة، ولا يسمح من ثم، بمزيد من اختبارات الفرضية. إن المستوى المهم للإدراج - في الوقت نفسه - يتغير باعتباره دلالة ثروة حالية، وعدداً للتكرارات منذ الرفض الأخير للفرضية الصفرية. وقد تم عرض هذا الإجراء لمراقبة احتمالية حالات الرفض الكاذبة للفرضية الصفرية بشكل فعال (Foster and Stine 2008).

إذن، إن انحدار معامل تضخم التباين، يشغل كُلّ متنبئ مرشح مرة واحدة فقط، وتعترف به إلا في حالة تجاوزه - إلى حدّ ما - شريط عالٍ للإدراج. ولكن، ألا يعني هذا أن الخوارزمية يمكن أن «تفقد» متنبئات مهمة؟ يؤكد لنا منتجو الخوارزمية،

عكس ذلك، إذ في حالة ما إذا كانت المتغيرات التنبؤية العالية غير مترابطة مع البقايا (Residuals)، فإن قاعدة استثمار ألفا ستضمن لنا أن النموذج برمته سيكون تنبؤياً، على الرغم من أنها لا تضمن دخول أي من المتغيرات، النموذج في حدّ ذاته. ويُنصح مستخدمو هذه التقنية بإيلاء الأولوية للمتغيرات الأكثر أهمية بعدهم أولاً. ويزعم مؤلفو البرنامج - ولغايات تنبؤية - عدم أهمية دخول متغير مترابط أو متغير آخر بشكل عالي، النموذج. وورد في كتاباتهم أنه «إذا كان بالإمكان تنظيف هذه الأهمية أو حجبها من لدن متغيرات أخرى، فسينعدم - ولغايات تنبؤية - أي فرق بين المتغير وبدائله، ومن ثم، عدم إمكانية اعتبار أي منها حالة «صادقة» (Lin, Foster, and Ungar 2011, 239). ولن يتم فقدان «متغيرات مهمة بشكل عام»، ولن يتم فقدان متنبئات إشارة عالية، في ظل انعدام متغيرات أخرى مترابطة بها ارتباطاً جوهرياً.

لقد تم إظهار خوارزمية انحدار معامل تضخم التباين، أسرع جوهرياً من الخوارزميات المنافسة (ويقترب طالب الطريق المُعَمم أكثر)، وأفضل في مراقبة معدل الاستكشاف الكاذب الهامشي (ولو أنها ليست جيدة مثل الانحدار التدريجي، أو الخوارزمية الأمامية - الراجعة [FoBa]، أو طالب الطريق المعمم)؛ فلديه أداء أفضل خارج العينة (دقة تنبؤية أكثر) من طالب الطريق المعمم، واللاسو، وهو جيد مثل «الفوبا» والانحدار التدريجي.

تشغيل انحدار معامل تضخم التباين: مثال باستخدام R

في حدود علمنا، إن الطريق الوحيد الذي يتم به تنفيذ انحدار معامل تضخم التباين، هو عبر حزمة R لمعامل تضخم التباين، الذي كتبه دونغيو لين (Dongyu Lin) (2011) ورسخه، أحد مطوري الطريقة. وننجز مثلًا من أمثلة انحدار معامل تضخم التباين مستخدمين مجموعة بياناتنا الخاصة بانتخابات 2012 على مستوى المحافظة. أما بخصوص مثالنا الذي يعرض «اللاسو»، فسننمذج حصة أوباما من الأصوات على مستوى المحافظة.

وسنقوم أولاً بتحميل الحزمة من الرابط: http://cran.r-project.org، وتثبيته في ذاكرة شغالة:

```
Install. packages ("VIF")
Library (VIF)
```

وبعد ذلك، نحوّل مجموعة من المتنبئات إلى مصفوفة، لأن معامل تضخم التباين يشتغل أفضل إذا ما تناولت X's باعتبارها شيئاً مستقلاً. وستكون إدارة (R's cbind) كافية لهذا الغرض. ولاحظ هنا أن لدى كُلّ المتغيرات حرف z باعتبارها سابقة (Prefix). ونقوم بذلك للدلالة على أن لدينا فارقاً معيارياً عن متوسط القيمة (z-score) لكُلّ المتنبئات التي نحن بصدد استعمالها. ولا بُدَّ من القيام بهذا لوضع كُلّ المتغيرات على مستوى واحد لكي يكون بالإمكان تقييم مساهمتها النسبية في تفسير التباين بشكل مناسب. وهذا ضروري بشكل عام في أي محدد سمة (Feature Selection) كي لا يكون هناك تحيز في اختيار الخوارزمية. ومع ذلك، بينما تقعِّدُ لك حزمة من قبيل بينلايزد، المتغيرات بمثابة خيار، يكون لزاماً عليك القيام به سلفاً - وعلى نحو سابق لأوانه - في ظل معامل تضخم التباين.

```
x<-cbind(zlnpop, zlnpopdens, zagelt18, zagel1834, zage3564, zagegt65, zperwhite, zperblack, zperamind, zperasian, zperpacisl, zperother, zpermultirace, zperlatin, zhhsize, zlthsed, zhsed, zsomecol, zbached, mastersed, zprofed, zdoced, zmalunemp, zmedinc, zperpov, zimdens, zprofmanocc, zdivorce2, zsamesex, zhigheredpop)
```

الآن وقد جمعنا كُلّ متنبئات مرشحنا ضمن مصفوفة، نحن مستعدين لآداء معامل تضخم التباين على النحو الآتي:

```
modl<-vif (zobama,x, w0 = 0.05, dw = 0.05, subsize = 200, trace = True)
```

ويولّد هذا شيئاً يدعى «مود 1» (mod 1)، سيتم داخله تخزين نتائج عملية انتقاء المتغير لمعامل تضخم التباين. وإن خيار w^0، تخبر البرنامج بالثروة الأولى التي نريد أن ينفقها النموذج. ومن أجل نموذج أكثر محافظة التي تنتقي متغيرات أقل، سنحدد هذه القيم في مستوى منخفض. وفي المقابل، إن تحديد الثروة الأولى أو تغيير في الثروة، في مستوى أعلى، سيسفر عن إدراج مزيد من المتغيرات. كما يخبر الحجم الفرعي (Subsize)، البرنامج بحجم العينة الفرعية العشوائية التي نحسب فيها معامل

تضخم التباين لكُلّ متغير على بساط البحث. وأ**خيراً**، إن «trace = TRUE» ، يمكننا بالاطّلاع على ما يقع عندما يدير معامل تضخم التباين مجموعة المتغيرات البالغ عددها الثلاثين، التي قدمناها من أجل التقييم. وإن القيام بذلك، يولّد المُخرج المبين في الشكل رقم 5.6.

ويمكن رؤية وجود 30 سطراً، واحد لكُلّ متغير من متغيرات المتنبئ 30، وخمسة أعمدة. وإن العدد الأول في العمود (بعد الرمز {1}) يخبرنا - ببساطة - عن المتغير الذي سيقيمه البرنامج. أما الأعمدة الأخرى فتخبرنا بما يلي:

1. الثروة الحالية (قبل تقييم المتغير الحالي)
2. مستوى الاختبار الحالي (الذي - تذكّر - يتغير مع كُلّ متغير جديد، استناداً إلى ما إن كانت المتغيرات القبلية قد وضعته ضمن النموذج أم لا).
3. إحصائية-t بالنسبة إلى المتغير قيد التقييم
4. قيمة-p بالنسبة إلى اختبار-t هذا.

وماذا يعني هذا كله؟ طيب، تأمل، ما يحدث في السطرين الأولين؛ ففي السطر الأول، لدينا الثروة التي اخترناها كنقطة انطلاق: 0.05. ولكي يتم إدراج المتغير في النموذج، فلا بُدَّ أن يكون ذا دلالة في $\alpha = 0.25$ (أو الثروة الحالية مقسومة على 2). إننا نرى أن نتيجة اختبار-t هي 22.77، وهي نتيجة أقل بكثير من $p<0.001$ (مقربة هنا إلى 0). وهذا يعني أن المتغير الأول الذي قدمناه لمعامل تضخم التباين، **zlnpop** (تقعيد-z، للخوارزمية الطبيعية للكثافة السكانية)، فسّرتْ تبايناً كافياً لإدراجه في النموذج. وفي السطر 2، نرى النتيجة: ارتفعت ثروة النموذج، في حين تم تحديد القيمة الحرجة (Critical Value) لإدراج المتغير الموالي في مستوى أقل انخفاضاً (في $\alpha = 0.18$) ومرة أخرى، إن إحصائية-t بالنسبة إلى المتغير الثاني، عالية جداً (6.953)، ويُقبَل المتغير في النموذج.

```
> mod1<-vif(zobama,X, w0=0.05, dw=0.05, subsize=200, trace=TRUE)
[1] "1 0.05 0.025 22.771554296814 0"
[1] "2 0.075 0.01875 6.95294747856659 3.57736062994718e-12"
[1] "3 0.10625 0.0177083333333333 5.72354728131594 1.04322517291422e-08"
[1] "4 0.138541666666667 0.0173177083333333 5.59771973911076 2.17189479734259e-08"
[1] "5 0.171223958333333 0.0171223958333333 4.83686222030236 1.31904790978687e-06"
[1] "6 0.2041015625 0.0170084635416667 0.828564198610133 0.40735105353865"
[1] "7 0.187093098958333 0.0133637927827381 27.0053667414444 0"
[1] "8 0.223729306175595 0.0139830816359747 3.80020276492099 0.000144577740259999"
[1] "9 0.25974622453962 0.0144303458077567 13.5027750931116 0"
[1] "10 0.295315878731864 0.0147657939365932 5.57551515371425 2.46798421699168e-08"
[1] "11 0.330550084795271 0.0150250038543305 0.793302156109217 0.427601800931797"
[1] "12 0.31552508094094 0.0131468783725392 0.759470464593518 0.44757117560337"
[1] "13 0.302378202568401 0.0116299308680154 5.24342826272885 1.57620108520717e-07"
[1] "14 0.340748271700385 0.0121695811321566 1.4465381095496 0.148026331031677"
[1] "15 0.328578690568229 0.010952623018941 10.5970808072016 0"
[1] "16 0.367626067549288 0.0114883146109152 11.6686643604467 0"
[1] "17 0.406137752938373 0.0119452280275992 5.68462572632097 1.31099446853966e-08"
[1] "18 0.444192524910773 0.0123386812475215 4.9769822834295 6.45832303414196e-07"
[1] "19 0.481853843663252 0.0126803643069277 6.14694924012578 7.89873944029296e-10"
[1] "20 0.519173479356324 0.0129793369839081 2.26986290639511 0.0232159023444118"
[1] "21 0.506194142372416 0.0120522414850575 0.498836920414769 0.617894275430784"
[1] "22 0.494141900887359 0.0112304977 4744 1.41894577497042 0.15591482535024"
[1] "23 0.482911403139919 0.0104980739813026 18.4175936639235 0"
[1] "24 0.522413329158616 0.0108836110241378 3.98169787457411 6.84247222204615e-05"
[1] "25 0.561529718134478 0.0112305943626896 0.531800210693972 0.594864376832877"
[1] "26 0.550299123771789 0.0105826754571498 0.570307523144667 0.568469139224964"
[1] "27 0.539716448314639 0.00999474904286369 6.86589616654364 6.60760335335908e-12"
[1] "28 0.579721699271776 0.0103521732012817 2.75214879766392 0.00592056128820428"
[1] "29 0.619369526070494 0.0106787849322499 2.24618282745469 0.0246922997211378"
[1] "30 0.608690741138244 0.0101448456856374 1.159582223814 0.246218941876399"
> |
```

الشكل رقم 5.6: النتيجة تُظهر انتقاء متغير من انحدار معامل تضخم التباين في R.

وفي المقابل، نستطيع رؤية ما يقع عندما يخفق متغير ما لجعله ضمن النموذج من خلال النظر إلى ما يقع قبل متغير 6، وبعده. ولاحظ أن ثروة النموذج ترتفع بالنسبة إلى كُلّ متغير من 1 إلى 6. وتذكر أن هذه هي ثروة النموذج قبل إخضاع المتغير الجديد إلى التجربة. والمتغير 6، لا يضعه في النموذج (التي نستطيع الإفصاح عنه من خلال النظر إلى قيمة-t 0.829، وقيمة-t 0.407). ولاحظ أنه بالنسبة إلى المتغير 7، تتراجع الثروة قليلاً (من 0.204 إلى 0.187). وبعد إدارة متغيراتنا، على ماذا سنحصل من حيث تناسبية النموذج؟ في الحقيقة، إن «روتين» معامل تضخم التباين، لا يتناسب مع النموذج بالنسبة إليك؛ بل على العكس من ذلك، إنه يخبرك عن المتغيرات التي يجب عليك ضمها، والمتغيرات التي يستوجب عليك رفضها. إنه - إذن - محدد سمة أصلي.

ولرؤية المتغيرات المختارة، نستعمل ما يلي:

```
mod1$select
```

وتعود R:

```
28 27 24 23 19 18 17 16 15 13 10 9 8 7 5 4 3 2 1 {1}
```

وهذا يخبرنا عن هوية أعداد المتغيرات - 19 في المجموع - التي اختارها النموذج. ونرى - من خلال فحص القائمة - أن العديد من المتغيرات مفقود:

```
Call:
lm(formula = zobama ~ X2)

Coefficients:
 (Intercept)      X2zlnpop   X2zlnpopdens      X2zagelt18       X2zage1834      X2zage3564
  -2.199e-09    -3.971e-02      2.688e-01      -2.985e-02        3.584e-02       1.275e-01
  X2zperwhite   X2zperblack    X2zperamind    X2zperasian  X2zpermultirace     X2zperlatin
  -2.354e+00    -1.306e+00     -4.710e-01      -1.571e-01       -2.679e-01      -1.171e+00
   X2zhhsize     X2zlthsed        X2zhsed      X2zsomecol        X2zbached     X2zmalunemp
  -1.752e-01    -7.381e-02     -3.069e-01      -3.513e-01       -2.283e-01       2.439e-01
   X2zmedinc  X2zprofmanocc    X2zdivorce2
  -1.008e-01    -1.479e-02     -4.882e-02
```

الشكل رقم 6.6: النتيجة من انحدار معامل تضخم التباين في R.

المتغيرات، 6، 11، 12، 14، 20، 21، 22، 25، 26، 29، 30. ومن المهم تذكر أن انتقاء متغيرات بواسطة انحدار معامل تضخم التباين يتوقف - إلى حدّ ما - على الترتيب المعتمد في إدراجها. إن معامل تضخم التباين يجرب كُلّ متغير مرة واحدة فقط، ويحاول ببساطة تعظيم القوة التفسيرية دون الإفراط في التناسب. ومن ثم، إذا كنا نحاول ضمان اختيار الخوارزمية للمتغيرات «الحقيقية»، فسيكون تشغيلها عدة مرات فكرة جيدة، من خلال تغيير ترتيب المتغيرات في كُلّ مرة.

وإذا ما قلصنا قيمة معلم dw، فسنقلص عدد المتغيرات المقبولة لدى النموذج، وذلك لأن القيمة الحرجة المدرجة في النموذج تتوقف على ثروة النموذج. وإن إضافة ثروة أقل إلى رفض فرضية صفرية ما، يؤدي إلى قيم حرجة من أجل إدراج أقل

انخفاضاً، ومن ثم، من أجل سمات منتقاة أقل. وعندما نحدد dw في 0.05، ينتقي معامل تضخم التباين 19 متغيراً. ولكن، هذا المَعلم، لا يفضي سريعاً إلى نموذج أكثر تقتيراً. وعندما نُخفّض من dw إلى 0.01، و0.0001، ننتقي متغيرات 18، و18، و17 على التوالي. أما البديل الآخر، فيتمثل في تقليص الثروة الأولية للنموذج. ولكن مرة أخرى، على المرء - بانتقاء المتغيرات هذه - تحديد w0 على نحو منخفض جداً قبل أن يصبح النموذج أصغر بكثير.

وبعد تسوية المتغيرات من أجل الإدراج نشغل - ببساطة - نموذجاً خطياً بالاستعانة فقط بتلك المتغيرات المنتقاة. ونشكل يدوياً مصفوفة، تضم فقط المجموعة الفرعية لمتغيرات منتقاة، وبعدها تشغيل نموذج خطي في هذه المجموعة الفرعية. ويظهر مُخرج انحدار R، في الشكل رقم 6.6.

```
X2<-cbind (zlnpop, zlnpopdens, zagelt18, zagelt1834, zage3564,
zperwhite, zperblack, zperamind, zperasian, zpermulti-race, zhhsize,
zlthsed, zhsed, zsomecol, zbached, zmalunemp, zmedinc, zprofmanocc,
zdivorce2)

Mod2<-lm(zobama~X2)
```

قد انتفى معامل تضخم التباين، المتغيرات التي سبق لنا أن لاحظنا أهميتها في تنبؤ حصة أوباما من الأصوات على مستوى المحافظة: الكثافة السكانية، نسبة السكان غير الإسبان من البيض، ونسبة السكان السود، ونسبة البالغين الشباب في السكان، ومعدل البطالة بين الرجال. وهكذا، تنتج هذه الخوارزمية الناجعة للغاية نتائج، تتوافق بشكل كبير مع نتائج النماذج التي رأيناها من قبل.

وكلما مضى توسع عالم التنقيب في البيانات قدماً على قدم وساق، مولّداً - إلى الأبد - خوارزميات جديدة، يكون الباحثون قد طوروا - مع ذلك - تقنية أخرى، تحسن ظاهرياً انحدار معامل تضخم التباين. وهذه الطريقة الحدثية - انحدار معامل تضخم التباين قوية - تعالج ميل انحدار معامل تضخم التباين «المعياري» لأن يصير حساساً لحضور حالات شاذة في البيانات (Dupuis and Victoria- Feser 2013). ويعد انحدار معامل تضخم التباين طريقة مهمة لانتقاء المتغيرات على نحو ناجع لتعظيم دقة تنبؤية.

الفصل السابع
إنتاج متغيرات جديدة

إن المتخصصين المتمرسين في التنقيب في البيانات، يخبرون الوافدين الجدد على الميدان باستمرار بأن ما يستغرق معظم الوقت عادة، ويتطلب العناية الكبرى في التنقيب في البيانات، ليس هو إدارة التحليل (مرحلة النمذجة)، بل هي المرحلة التي تسبق تحليل البيانات عندما ينتج الباحث المتغيرات أو السمات التي ستدخل ضمن نماذج. ويرجع سبب ذلك - جزئياً - إلى استخدام الباحثين معرفتهم بالموضوع لضمان عدم إهمال متغيرات هامة. كما يعمل الباحثون أيضاً على تشكيل النسب التي تبدو هامة من حيث التصور (التكلفة للقدم المربع الواحد، عمليات إطلاق النار بحسب 100,000 نسمة، وهكذا)، وقد تظهر متنبئات قوية تجريبياً. وفوق هذا، مهما يدرك متخصصو التنقيب في البيانات إمكانية أن يكون شكل المتغيرات مهم بالنسبة إلى التحليلات التي تلي، فإن عليهم الأخذ بعين الاعتبار تحولات ممكنة لمتغيراتهم.

وتهم الحالة الأكثر بساطة وشيوعاً، التقعيد أو المعيارية (Standardization). وفي بعض الطرق - لكن ليس كُلّ الطرق - ستفضل الخوارزمية، المتغيرات التي تمتلك فئات عديدة (أو مجموعة كبيرة من القيم)، مثل العمر بالسنوات، أو الدخل بالدولارات، باعتبارها أكثر تنبؤاً أو ترابطاً - منطقياً - من متغير ذي فئات أقل، أو ذي مجموعة أصغر من القيم، من قبيل الحالة الاجتماعية. ونقصد بكلمة «تُفضل» أن البرنامج سيعتبر متنبئاً ما، ذا فئات عديدة، أو مجال واسع، متنبئاً أكثر قوة من متنبئ لا

يتوافر إلا على فئات أقل. وينبع هذا التحيز من الطريقة التي نقدم بها بياناتنا، عوض عكس البنية الحقيقية في البيانات. على سبيل المثال، قد نختار تمثيل الدخل بالدولارات، أو بدولارات مسجلة. وقد نمثل العمر بالسنوات أو نمثله ضمن مجموعات مثل المراهقين (العشرات)، والعشرينات، والثلاثينات، وهكذا. وستتغير علاقة المتغير بالنتيجة بحسب نوع الاختيارات التي نتخذها، مخلفة أحياناً تأثيراً في الأهمية التنبؤية الظاهرية لهذا المتغير المرتبط بمتغيرات أخرى. ويكمن الحلّ في تحويل كُلّ المتنبئات المترشحة في مجموعة بيانات لشخص ما، داخل مقياس مشترك، وهي عملية معروفة بالتقعيد.

إن نوع التقعيد الأكثر شيوعاً، يحول المتغيرات المستمرة (سواء قيست باعتبارها متغيرات فاصل زمني/ نسبة أو متغيرات ذات مستوى عادي) إلى فوارق معيارية عن متوسط القيمة (z-scores)، ولهذه الفوارق المعيارية متوسط الصفر، وانحراف معياري قيمته واحد. ولهذا، مهما بلغ الفرق في محتواها (العمر، والدخل، ومعدل الذكاء، وساعات العمل بالأسبوع)، بعدما تم تقعيد معدلاتها، فسيكون لدى المتغيرات المحولة المعدل أو المتوسط نفسه، والانتشار نفسه.

ويحدث نوع آخر من أنواع إنتاج المتغير، عندما يأخذ المرء متغيراً مستمراً، مثل العمر بالسنوات، والدخل بالدولارات، ويحوّله أو يغير تشفيره إلى مجموعة فئات مميزة ومنظمة، على سبيل المثال، إنتاج فئات عمرية مثل 0-10، 11-20، 21-30، وهكذا، إلى أن تصل إلى أعمار تتراوح بين 71-80. ويعرف هذا «التقطيع» للمتغيرات المستمرة بالتنقيب في البيانات باعتباره عملية توزيع خانات (Binning)، أو تفريداً أو تميزاً (Discretization) وإلى جانب إنتاجه للنسب والمتنبئات التي تم تقعيد معدلاتها، فإن هذا التقطيع يعد الخطوات الأكثر انتشاراً في عملية معالجة البيانات مسبقاً قبل تشغيل نموذج ما.

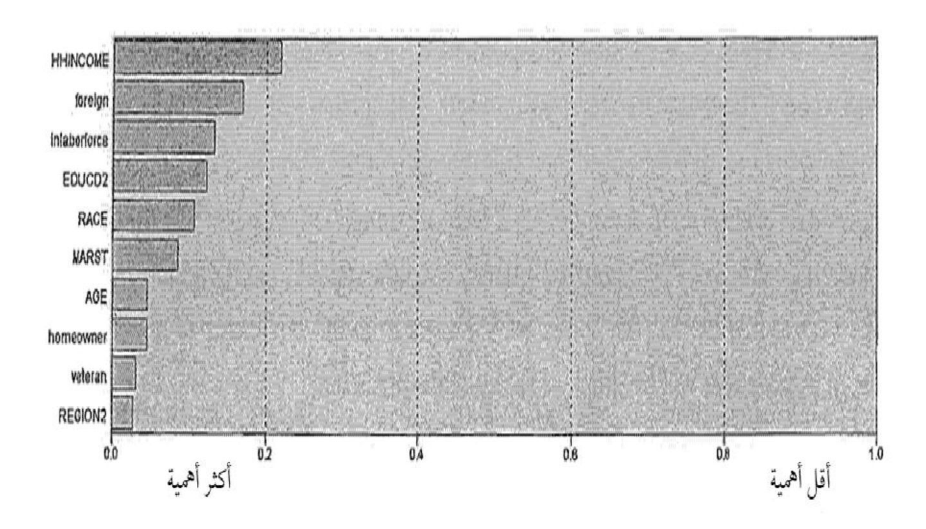
وحسب الانطباع الأول، يبدو أن تغيير متغير مستمر ما، مثل العمر بالسنوات، إلى متغير متميز مثل فئة عمرية، يفترض ضمناً فقدان المعلومات، أي عدم وضوح التفاصيل الدقيقة. هذا صحيح، ولكن توزيع الخانات لها ميزة التعويض التي تجعل من السهل جداً تمييز العلاقات اللا خطية (ونمذجتها) بين متنبئ ومتغير النتيجة. لنقدم مثالاً عن ذلك.

سنستعمل البيانات المأخوذة من مسح المجتمع الأميركي للتنبؤ فيما إذا كان شخص ما يمتلك تأمين صحي، مستخدمين خصائص ديموغرافية متعددة. ونشكل مجموعة بياناتنا الخاصة لاحتواء عدد مساوٍ من حالات الأفراد المؤمّنين وغير المؤمّنين. ولأن الأشخاص غير المؤمّنين يشكلون أقلية صغيرة نسبياً ما، السكّان (حوالي 14٪)، سنبقي على جميع الحالات غير المؤمّنة في العينة، على أن تكون عينة عشوائية مأخوذة من أكبر عدد من الأفراد المؤمّنين، بغية الحصول على تقسيم 50:50. وإننا نقوم بذلك لأنه في حضور تصنيف نتائج غير متوازنة بشكل كبير، تعمل الخوارزميات - في الغالب - على تصنيف كُلّ الحالات باعتبارها حالات تنتمي إلى فئة أغلبية لتقليص معدل الخطأ في التنبؤ. وتؤدي الموازنة في البيانات إلى نموذج أهم، واختبار أفضل من دقة تنبؤية.

ندير انحداراً لوجيستياً يتنبأ بوضعية التأمين الصحي، بحيث يحمل ترميز 1 إذا ما كان يفتقر فرد ما إلى تأمين صحي، وترميز 0 إذا ما كان مؤَمَّناً. ويدخل ضمن المتنبئات كُلّ من الدخل، والعمر، والحالة الاجتماعية، والعرق، والجنوسة (Gender)، وملكية المنزل، والتحصيل التربوي، ومكان الميلاد، والخدمة العسكرية، وحالة القوة العاملة، ومنطقة التعداد. ومن المهم هنا تأكيد دخول العمر والدخل ضمن النموذج باعتبارهما متغيرين مستمريين. ولدى نموذجنا شبه مربع مكفادين (McFadden's Pseudo-R^2 0.202)، وتسجيل احتمال (Log-Likelihood) يصل إلى 565007.947-. ويبين الجدول رقم 1.7 مصفوفة الارتباك، (التي تجدول الفئة المتوقعة، مقابل الفئة الحقيقية للنتيجة)، في حين يمثل الشكل رقم 7.1 تصوراً للأهمية النسبية لمتغيرات مستقلة بالنسبة للتنبؤ. وكلاهما يمكن توليدهما بشكل آلي بواسطة الحزمة الإحصائية للعلوم الاجتماعية (SPSS) بعد تشغيل انحدار لوجيستي.

الجدول رقم 1.7: مصفوفة الارتباك للانحدار اللوجيستي باستخدام بيانات متوازنة (الحزمة الإحصائية للعلوم الاجتماعية، SPSS).

	مؤمن متنبأ	مؤمن غير متنبأ	دقة
مؤمن حقيقي	78,974	32,409	٪70.90
غير مؤمن حقيقي	31,242	79,788	٪71.86



الشكل رقم 1.7: أهمية المتنبأ بالنسبة إلى الانحدار اللوجيستي باستخدام بيانات متوازنة (الحزمة الإحصائية للعلوم الاجتماعية، SPSS). الهدف: غير مؤمن.

إن معدل الدقة بشكل عام هو 71.4٪، إذ يعمل بحق، وعلى نحو مماثل، على تصنيف إيجابيات صادقة، وسلبيات صادقة على مستوى النتيجة. وبحسب نتائجنا، يمثل دخل الأسرة، المتغير الأكثر تنبؤاً في نموذجنا، متبوعاً بمكان الولادة (متغير وهمي(Dummy Variable) بالنسبة إلى الأفراد المولودين في الخارج)، وحالة القوى العاملة، والتحصيل التربوي. ولا يبدو العمر مهماً - نسبياً - في تنبؤ تغطية التأمين الصحي؛ بل هو معارض للحدس، بما أننا ندرك بأن الحقّ في بعض برامج التأمين الصحي (أبرزها الرعاية الصحية) يقوم أساساً على العمر. كما أن المساعدة الطبية للفقراء، وبرنامج الدولة للتأمين الصحي للأطفال (SCHIP)، متاحة أيضاً للأفراد - جزئياً - على أساس العمر (وجزئياً على أساس الدخل). وحسب مخرج انحدارنا (غير مبين)، يملك العمر علاقة سلبية معتدلة مع حالة التأمين الصحي. وربما، يتجلى المشكل في عدم خطية العلاقة بين العمر وحالة التأمين الصحي. وربما ثمة احتمالات مختلفة بشكل مميز للتأمين لفائدة أشخاص في مجموعات عمرية مختلفة.

تفريد متنبئ مستمر

إن أشجار التقسيم - المعروفة أيضاً بأشجار القرار أو أشجار التصنيف - تعد خوارزميات تنبؤية، تستكشف الطريقة الأكثر نجاعة التي من خلالها يتم فصل الحالات بالنسبة إلى فئات نتيجة ما. وسيتم وصفها بتفصيل في الفصل العاشر، غير أننا سنركز حالياً على كيفية إمكانية استخدامها لوضع متغيرات مستمرة في خانة أو تمييزها من أجل تحسين تنبؤ نتيجة ما.

ولتصنيف الحالات، تقسّم أشجار التقسيم حالات حسب كُلّ قيمة لكُلّ متغيرات التنبؤ المرشحة، والمحددة من قبل الباحث؛ فتجد ذلك التقسيم الذي يفصل - بشكل أفضل - الحالات إلى فئات النتيجة قيد البحث. وبإيجادها هذا التقسيم المثالي، تواصل إعادة هذا الإجراء إلى أن تنتج مجموعات متجانسة من حيث النتيجة، أو إلى أن يصدر الباحث تعليماته بإيقاف البرنامج.

عموماً، تأخذ أشجار التقسيم عدداً كبيراً من متغيرات مترشحة لدى اختيارها مكان التقسيم، ولكن يبقى عدد المتغيرات المستقلة التي تستخدمها الخوارزمية من صلاحية الباحث. وبعد ذلك، سيقسّم البرنامج فقط على هذا المتغير. وسيجد هذا - في الواقع - النقاط الفاصلة (Breakpoints) في المتغير على مستوى علاقته بالنتيجة. ومن ثم، إذا وجدت لا خطيات (Nonlinearities) معقدة بين متنبئ مستمر، ومتغير نتيجة ثنائي، فستكون أشجار التقسيم طرقاً ممتازة لإيجادها. وتعد «الأشجار» التي تنتج عن تطبيق شجر التقسيم على هذا النحو، خانات لقيم المتنبئ.

في مثالنا، نحاول تنبؤ عدم وجود تغطية التأمين الصحي، ونشك في إمكانية أن يكون نسبة العمر تنبؤية بشكل كبير، بل نرى أن العلاقة بين العمر والحالة الصحية غير خطية. ومن ثم، نستعمل شكلًا خاصاً من أشكال شجرة التصنيف المعروفة مربع كاي للكشف عن التفاعل التلقائي (CHAID)، المشغل في إحصائية الحزمة الإحصائية للعلوم الاجتماعية (SPSS)، لفحص هذه العلاقة. وتوجد النتائج ملخصة في الجدول رقم 2.7.

وتقترح الشجرة طريقة لخلق مجموعات عمرية مثالية بالنسبة إلى تنبؤ حالة

التأمين الصحي، بحيث يضم الصنف العمري الأول الذي تم إنتاجه، أشخاصاً تتراوح أعمارهم بين 0-8، في حين تضم الخانة الثانية أشخاصاً تتراوح أعمارهم بين 9-17، وأما الخانة الثالثة، فتضم أشخاصاً تتراوح أعمارهم بين 18-24، وهكذا. وإن فحص مدى تغيير احتمال عدم كون الشخص مؤمَّناً صحياً عبر هذه الفئات العمرية، يخبرنا بمدى لا خطية العلاقة بين العمر وحالة التأمين. وفي هذه البيانات (التي - تذكّر - تمت موازنتها من حيث النتيجة)، يكون الاحتمال النسبي لكون الشخص غير مؤمن، منخفضة لدى الأطفال[1]؛ بينما يرتفع بشكل كبير بين الشباب الذين تتراوح أعمارهم بين 18-30. إننا نشهد - إذن - انحداراً بطيئاً في هذا الاحتمال، يشمل باقي مرحلة البلوغ. وفي المجموعة العمرية الكبرى (الذي حددته الحزمة الإحصائية للعلوم الاجتماعية في 63 عاماً)، يتراجع احتمال نسبة الأشخاص غير المؤمّنين.

وباختصار، إن العلاقة الحقيقية بين العمر واحتمال عدم كون الشخص مؤمناً، علاقة لا خطية، أي ترتفع وتنخفض عبر الطيف العمري. وفي السابق، لما أدخلنا العمر باعتباره متغيراً مستمراً في انحدارنا اللوجيستي، لم نستطع وضع اليد إلا على علاقة متوسط هامشي بين العمر وحالة التأمين الصحي، الذي كان عاجزاً بشكل مطلق عن رسم خريطة هذا التعقيد. ونتيجة لذلك، كان يبدو العمر غير مهم نسبياً في تنبؤ حالة التأمين. وكان ذلك - باختصار - نتيجة خطأ مواصفة (Specification Error). علاوة على ذلك، بما أن العلاقة بين العمر وحالة التأمين تحركها اقتطاعات قانونية عشوائية من أجل أهلية البرنامج، فإن نمذجة هذه العلاقة - ببساطة - بشروط تربيعية (Quadratic)، أو تكعيبية (Cubic) بالنسبة إلى العمر، لا تبدو أنها مرضية تماماً (ولو أنها ستشكل - بكل تأكيد - تطوراً مقارنة بالمواصفة الخطية). وسنبين أدناه، كيف أن عملية توزيع العمر في خانة بشكل مثالي ضمن فئات، وضم هذه الفئات باعتبارها متغيرات وهمية، يمكن أن يحسن القدرة التنبؤية للنموذج.

(1) وبما أننا وازنّا هذه البيانات في النتيجة، فإن الاحتمال الشرطي للحصول على التأمين - مع الأخذ بعين الاعتبار العمر الملخص في الجدول رقم 2.7 لا يتوافق مع الكميات الحقيقية للساكنة. ولكن بما أن المجموعات المؤمّنة، وغير المؤمّنة، تمت معاينتها عشوائياً (بمعدلات مختلفة)، فإن الفوارق النسبية في الاحتمال بين المجموعات العمرية، تعد مفيدة.(المترجم)

عقدة	مجموعة عمرية	٪ بدون تأمين صحي	عدد الحالات
1	0–8	30.73	49,685
2	9–17	40.43	54,745
3	18– 23	75.97	55,077
4	24–29	70.71	52,535
5	30–35	63.51	45,247
6	36–42	59.43	54,339
7	43–48	56.08	50,676
8	49–55	51.51	57,490
9	56–62	44.43	46,305
10	63+	7.87	54,548

الجدول رقم 2.7: استخدام شجرة مربع كاي للكشف عن التفاعل التلقائي (CHAID) لوضع متغير مستمر (عمر) في الحزمة الإحصائية للعلوم الاجتماعية، بشكل مثالي.

ولكن أولاً، نعود – في الجدول 3.7 إلى تحليل شجرة لعلاقة أخرى بين متغير مستمر – العائد الأسري – والتأمين الصحي. ومرة أخرى، تمكن البرنامج من تحديد النقاط الفاصلة في المتغير المستمر من حيث علاقته بالنتيجة؛ وتبدو هذه النقاط الفاصلة في 14,596$، و23,000$، و40,000$، 31,200$، وهكذا. ولكن يشير التفتيش حول كيفية تغير حالة التأمين عبر هذه المجموعات ذات الدخل، إلى علاقة خطية (أو على الأقل علاقة رتيبة) بين العائد الأسري والتأمين. وفي المجموعتين ذات الدخل المتدني، تفتقر نسبة كبيرة – نسبياً – من الأفراد إلى تأمين صحي. وتنخفض هذه النسبة كلما اتّجهنا تصاعدياً على مستوى الدخل إلى أن نحصل على الفئة ذات الدخل العالي جداً. وإن احتمال عدم توافر التأمين لفائدة هذه المجموعة، يمثل ثلث تلك المجموعة التي يعيش أفرادها على الدخل المتدني. ومؤدى ذلك أننا من غير المرجح الحصول – بشكل كبير – على نتيجة، على مستوى القوة التنبؤية من خلال استبدال مواصفة مستمرة للدخل بفئات ذات مجموعة الدخل (وإن كان علينا

البحث في هذا على كُلّ حال)، وفي الواقع، من المرجح فقدان القوة التنبؤية. وتتجلى المسألة هنا، في عدم استخدام الخانة دون تمييز. وفي حالات تكون فيها اللا خطية المعقدة ميزة من ميزات العلاقة «الحقيقية» الكامنة بين متنبئ مستمر ونتيجة ما، ستساعد على التنبؤ. ولكن إذا كانت العلاقة الكامنة خطية بكل تأكيد، لن تكون مساعدة، وستكون - في واقع الأمر - غير مناسبة.

مثال خلال استخدام إحصائية الحزمة الإحصائية للعلوم الاجتماعية

لقد بينا فقط كيف يمكن استخدام الأشجار لعملية توزيع المتغيرات المستمرة في خانة، ولكن على القراء أن يكونوا على علم بأن رزم متعددة من رزم برمجيات التنقيب في البيانات، تقدم تطبيقات تستطيع توزيع متغيرات مستمرة في خانة بشكل مباشر أكثر، دون أن يكون المستخدم مرغماً على فحص برنامج شجرة ما وتفسيره (ولو أن الرياضيات الكامنة، شبيهة جداً بتلك التي تعمل في الأشجار). كما تمكّن هذه الرزم المستخدم - بشكل آلي - بخلق وحفظ المتغير الجديد الموزعة في خانة أو المميز، في مجموعة البيانات، وهو أمر مريح. ونبين ذلك من خلال استخدام إحصائية الحزمة الإحصائية للعلوم الاجتماعية 21 لخلق 9 خانات من المتغير بالنسبة إلى العمر، ليتم التركيز مرة أخرى على حالة عدم التأمين الصحي باعتبارها نتيجتنا التي تهتم بها.

الجدول رقم 3.7: استخدام شجرة مربع كاي للكشف عن التفاعل التلقائي (CHAID) لوضع الدخل في الحزمة الإحصائية للعلوم الاجتماعية، بشكل مثالي.

عقدة	الدخل الأسري	٪ بدون تأمين صحي	عدد الحالات
1	14,595 أو أقل	63.81	52,014
2	14,596 - 23,000	65.18	52,163
3	23,001 - 31,200	63.65	52,259
4	31,201-40,000	61.20	53,355 5
5	40,001-49,997	56.38	48,617
6	49,998-61,400	51.59	53,983

7	61,401-76,000	45.35	52,216
8	76,001-96,990	38.30	51,713
9	96,991-133,500	31.33	52,257
10	أزيد من 133,500	23.17	52,050

صيغة عملية التمييز المثالي هي:

```
OPTIMAL BINNING
/VARIABLES GUIDE = uninsured BIN = AGE SAVE = NO
/CRITERIA METHOD = MDLP PREPROCESS = EQUALFREQ
    (BINS = 9)
FORCEMERGE = 0 LOWERLIMIT = INCLUSIVE
LOWEREND = UNBOUNDED UPPEREND = UNBOUNDED
/MISSING SCOPE = PAIRWISE
/PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

وكما هو مبين في الجدول رقم 4.7، إن الحزمة الإحصائية للعلوم الاجتماعية تنتج تسع خانات بالنسبة إلى العمر. والمجموعات العمرية التي تم إنتاجها هنا شبيهة بتلك التي أنتجتها شجرة مربع كاي للكشف عن التفاعل التلقائي أعلاه. ولم يعد الشباب مقسمين إلى مجموعتين، ولكن ظهروا باعتبارهم مجموعة واحدة تتراوح أعمارهم بين 18-30. والنقطة الفاصلة بين المجموعة الأكبر سناً هي الآن 64 عوض 63 (ومع ذلك ليست 65، وهو الأمر غير المتوقع إلى حدّ ما).

نود الإشارة إلى مسألة أنه لو حددنا متغيراً مستقلاً مختلفاً، فسيكون من المرجح أن يولّد البرنامج مجموعات عمرية مختلفة. إن العملية المثالية لعملية توزيع الخانات ينتج فئات من متغيرات مستمرة مثالية من حيث تنبؤ نتيجة معينة. ومهم تذكر أن العملية المثالية لتوزيع الخانات ليست أمراً تم القيام به في بداية مشروع التنقيب في البيانات، مع إمكانية استعمال فئاته المحصل عليها في تنبؤ العديد من المتغيرات التابعة المختلفة. وكل عملية من العمليات المثالية خاصة بمتغير تابع أو متغير نتيجة واحد.

الجدول رقم 4.7: إنتاج فئات عمرية بواسطة عملية توزيع خانات مثالية (إحصائية الحزمة الإحصائية للعلوم الاجتماعية، SPSS).

العمر					
الخانة	نقطة نهاية		عدد حالات حسب مستوى غير مؤمّن		
	أدنى	أعلى	**0.00**	**1.00**	مجموع
1	غير مقيد	10	48,967	21,736	70,703
2	10	18	42,109	26,134	68,243
3	18	30	41,249	109,838	151,087
4	30	37	27,523	49,651	77,174
5	37	44	31,714	46,119	77,833
6	44	50	31,768	40,588	72,356
7	50	56	34,494	37,071	71,565
8	56	64	42,340	34,610	76,950
9	64	غير مقيد	71,667	6,084	77.751
المجموع		371.831	371.831	743,622	

ملاحظة: كل جزئية احتسبت على أساس أنها أدنى ≤ العمر ≤ الأعلى (lower ≤ Age ≤ Upper)

أما ولدينا الآن العمر في خانات مميزة، يمكننا إدارة انحدار لوجيستي جديد لمعرفة ما، إن حسّنت عملية وضع الخانات النموذج. وكما تم ذلك في السابق، نقدم مصفوفة الارتباك (الجدول رقم 5.7) ورسم بياني لـ أهمية المتنبئ (الشكل الرقم 2.7). وإن نموذج انحدارنا اللوجيستي الجديد له شبه مربع مكفادين (McFadden's

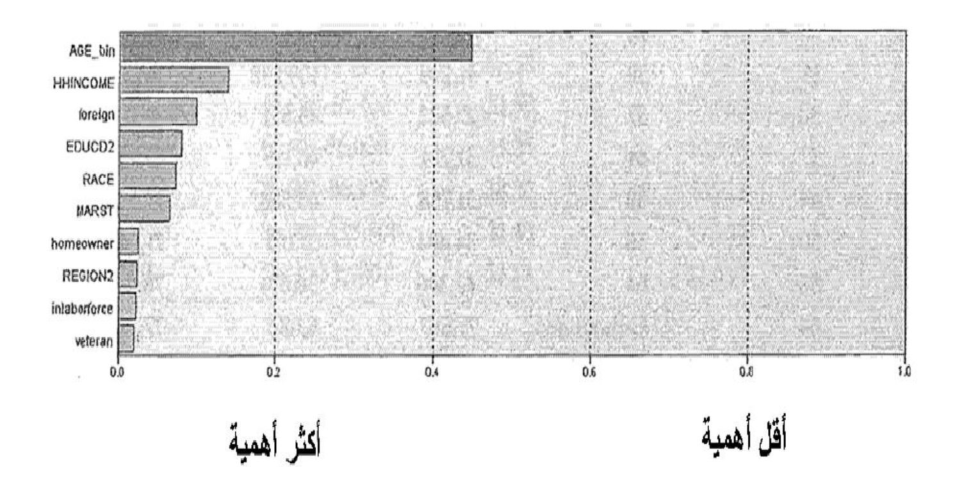
($pseudo-R^2$)، 0.259، وتسجيل احتمال (Log-Likelihood) يصل إلى 486350.891، وكلاهما يشير إلى نموذج ذي تركيبات أفضل (Better-Fitting). وتوضح مصفوفة الارتباك بتحقيقنا بعض التطورات في تصنيف كُلّ من الإيجابيات والسلبيات الصادقة.

والشيء الأهم من كُلّ هذا، هو أننا نلاحظ في الشكل رقم 2.7 (مقارنة مع الشكل رقم 1.7)، أن العمر حتى الآن، المتنبئ الأهم لحالة التأمين الصحي؛ يعد الآن أكثر أهمية مرتين من الدخل الأسري. وإن ترتيب المتغيرات من حيث الأهمية لا يتغير بشكل كبير، مما يقترح أن العمر لا يصف الآن تباينا من التباينات السابقة التي تم وصفها سابقاً بواسطة متغيرات أخرى. لقد نتج عن عملية توزيع العمر في خانات، تحسناً حقيقياً في النموذج، عوض إعادة توزيع (Reallocation) «العمل» انطلاقاً من متغيرات أخرى إلى العمر.

لقد بيّنا في هذا القسم كيف يمكن استخدام أشجار التصنيف، وعملية توزيع الخانات المثالية، استخداماً مثمراً لاستكشاف اللا خطية في العلاقة بين متغيرات المتنبئ المستمرة وبين متغير نتيجة ثنائية التفرع. كما رأينا أيضاً كيف أن عملية استكشاف هذه العلاقات اللا خطية يمكن أن تفرز تحسناً في القدرة التنبؤية. بعد ذلك، نعود إلى العلاقة بين متنبئات مستمرة، ونتائج مستمرة، ونوضح كيف أن ممارسات مماثلة يمكنها أيضاً أن تكون مفيدة في هذه الحالة.

الجدول رقم 5.7: مصفوفة الارتباك المتنبئة للتأمين الصحي مع تمييز العمر.

	مؤمّن متنبأ	غير مؤمّن متنبأ	دقة
مؤمّن حقيقي	82,914	29,181	٪ 73,97
غير مؤمّن حقيقي	27,099	84,746	٪ 75,77



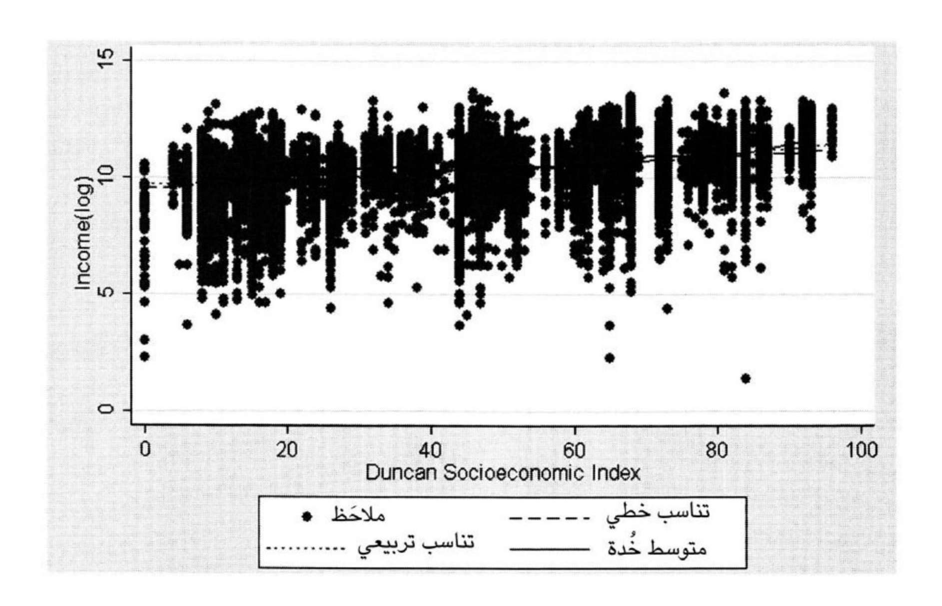
الشكل رقم 2.7: أهمية المتنبئ باستخدام المتغير العمري ذي الخانة في إحصائية الحزمة الإحصائية للعلوم الاجتماعية. الهدف: غير مؤمّن.

نتائج مستمرة ومتنبئات مستمرة

إن المنطق نفسه الذي تم العمل به في حالة من حالات متغير نتيجة ثنائية التفرع، يمكن تطبيقه أيضاً على النتائج المستمرة. دعنا نقول إننا نحاول تنبؤ نتائج مقاسة بشكل مستمر مثل الدخل، مستخدمين سمة مقاسة أخرى بشكل مستمر. وإذا كانت العلاقة بين المتغيرين خطية، فإن الانحدار الخطي المعياري يمكن استخدامه بسهولة. وإذا كانت العلاقة منحنية الأضلاع (Curvilinear)، فإمكاننا إضافة قيم تربيعية، أو تكعيبية، أو قيم ذات ترتيب أعلى لتقريب العلاقة. وإذا وجدت نقطة أو مزيد من النقاط الفاصلة الواضحة، فبإمكاننا نمذجة العلاقة بشكل جيد من خلال مواءمة خُدّة (Spline) ما.

ولكن أحياناً، يمكن أن يرتبط متغيران اثنان على نحو أكثر تعقيداً. على سبيل المثال، لندرس العلاقة بين المكانة المهنية (Occupational Prestige)، والدخل (مسجل)، في الشكل رقم 3.7. على الرغم من وجود اتّجاه تصاعدي عام في الدخل، كلما اتّجهنا نحو قيم أعلى من المكانة المهنية، فمن الواضح وجود فواصل وانقطاعات في هذه العلاقة، ليست مضبوطة بشكل جيد بواسطة نمذجة خطية. وفي

الشكل رقم 3.7، قمنا بضم خط تربيعي تنبؤي في الترسيمة (Plot)، ولكنه يلي التنبؤ الخطي بشكل كبير.



الشكل رقم 3.7: خطوط مخطط التشتت (Scatterplot) والمناسبة الواصفة للعلاقة بين الدخل الشخصي والمكانة المهنية في مسح المجتمع الأميركي (مأخوذ عن «الستاتا» Stata).

وإن عملية إضافة قيم ذات ترتيب عالي لا يعزز القوة الإيضاحية بقدر كبير، وهو أمر يتأكد من خلال أداء انحدار ما. (النتائج في الجدول رقم 6.7). إن القيمة الخطية بالنسبة إلى المكانة المهنية بمفردها تفسر حوالي 14٪ من التباين في دخل مسجل. وإن إضافة قيمة تربيعية، وبعدها قيمة تكعيبية، يعزز قوة إيضاحية بأقل من 1 في المائة نقطة. (إن المعاملات بالنسبة لهذه القيم ذات الترتيب العالي، تحقق دلالة إحصائية (Statistical Significance) في $p<.001$، ولكن يتم ذلك قبل كُلّ شيء لأننا بصدد استخدام مجموعة بيانات تفوق 340.000 حالة. ومن خلال هذه القوة الإحصائية الكبيرة، سيكون - عملياً - كُلّ متغير ذي دلالة إحصائياً في مستويات «ألفا» المعيارية).

ربما يتجلى المشكل - ببساطة - في وجود نقاط انعطاف (Inflection Points) كافية في نموذج تكعيبي، وفي عملية مواءمة خُدٍّ ما، قد تكون أكثر ملاءمة. وفي الشكل رقم 3.7، نوضح أيضاً متوسط خدة مناسبة، يضم عُقَداً متباعدة بشكل متساوٍ، ولكن الخدة لا تختلف في الطريق كثيراً عن التنبؤ الخطي. كما أن إضافة قيم ذات قوة أعلى لهذا النموذج، لا يعزز القوة التنبؤية، لأن العلاقة أكثر تعقيداً، مما يسمح به هذا النموذج. وإن إضافة الخدة لا يساعد، لأننا لا ندرك عدد نقاط الانعطاف في العلاقة بين المتغيرات، ولأن نقاط الانعطاف تلك ليست - على ما يبدو - متباعدة بشكل متساوٍ. وفي هذه الحالة، يمكننا الاستفادة من تقنيات التنقيب في البيانات كي تساعدنا على العمل بشكل أفضل.

وعندما تكون العلاقة بين متغيرين معقدة على النحو الذي نراه هنا، نستطيع نمذجته بإنتاجية أكثر من خلال تقسيم بياناتنا إلى خانات منفصلة للمتغير الإيضاحي، وبعدها استخدام مجموعة من المتغيرات الوهمية لهذه الخانات. ولكن في حدود أي قيم من قيم مؤشر دونكا السوسيو اقتصادي (Duncan Socieconomic Index)، يتوجب علينا القيام بتقطيعاتنا؟

نستخدم دالة تقسيم «غامب برو» للقيام بذلك بالنسبة إلينا، ونستخدم خيار الصلاحية المتبادلة لمطوية (K-Fold) (بثلاث طيات). وسيمكّننا هذا من الفصل فيما إن كنا بصدد الإفراط في عملية تناسبية النموذج. ولكننا، نستعمل 10٪ من عينة مسح المجتمع الأميركي لعام 2010، الذي يضم حوالي 340,000 حالة. وبهذه الحالات المتعددة، يمكننا بناء نموذج معقد جداً من دون إفراط في التدريب.

وإذا سمحنا بتشغيل النموذج حيث بداية صلاحية R^2 في التراجع، ستقسم الشجرة البيانات إلى 79 مرة، مشكلة بذلك خانات المؤشر السوسيو اقتصادي. الآن، توجد في هذه البيانات فقط 81 قيمة من قيم المؤشر السوسيو اقتصادي المتميزة. وهذا يعني أن البرنامج أنتج خانة منفصلة بالنسبة إلى كُلّ قيمة منفصلة على حدة. وهذه النتيجة - مع ذلك - هي دالة لكُلّ من عملية منح أولوية للتنبؤ على إمكانية التفسير (Interpretability)، وللحجم الكبير جداً لبياناتنا. وتوضح مجموعات البيانات الضخمة عملية المقايضة بين إمكانية التفسير والدقة التنبؤية، بطريقة لا تقدر

عليها مجموعات البيانات الضخمة التقليدية. إننا نريد أن نبسط نموذجنا الخاص بالبيانات، بطريقة مفيدة - كي نختصر ذلك. لكن ببساطة، لا يوجد قدر كبير من المقايضة في مجموعات بيانات ضخمة بين التعقّد (Complexity) والدقة. وسنكون في حاجة إلى فرض قيد على التعقّد إلى درجة تبدأ فيها إمكانية التفسير في الانحدار. وسيوضع هذا القيد بشكل عشوائي جداً، ومن ثم نقرر إعادة شذب الشجر للحصول على 12 تقسيمات. وعند هذا العدد من التقسيمات، نكون قد ضحينا فقط بقدر صغير من الدقة التنبؤية، ولكننا في الوقت ذاته، نكون قد حسنا من إمكانية التفسير.

الجدول رقم 6.7: نماذج انحدار المربعات الصغرى العادية (OLS) المتنبئة (لسجيل) الدخل من خلال مؤشر دونكا السوسيو اقتصادي.

	نموذج 1	نموذج 2	نموذج 3
المؤشر السوسيو اقتصادي	0.0169***	0.0061***	0.0310***
المؤشر السوسيو اقتصادي^2		0.0001***	−0.0004***
المؤشر السوسيو الاقتصادي^3			<0.0001***
ثابتة	9.523	9.692	9.455
R^2	0.143	0.147	0.149
جذر متوسط مربع الانحراف (RMSE)	0.998	0.996	0.995

المصدر: مسح المجتمع الأميركي، 2010.
$***p < .001$.

إن حلّ المجموعات الثلاثة عشر التي قمنا بتسويتها، يمكن ملاحظته في الشكل رقم 4.7، والجدول رقم 7.7 وكما رأينا سابقاً، إن العلاقة العامة بين المكانة المهنية والدخل إيجابي، ولكن النمو ليس رتيباً. وفي الثلثين الأقل انخفاضاً من معدلات قياسات المؤشر، يوجد نمط من الزيادات والانخفاضات في الدخل، وإيحائية المقايضات بين الدخل والحالة الاجتماعية. ويمكن لهذه العلاقة المميزة بلا خطية معقدة، ملاحظتها في الشكل رقم 5.7.

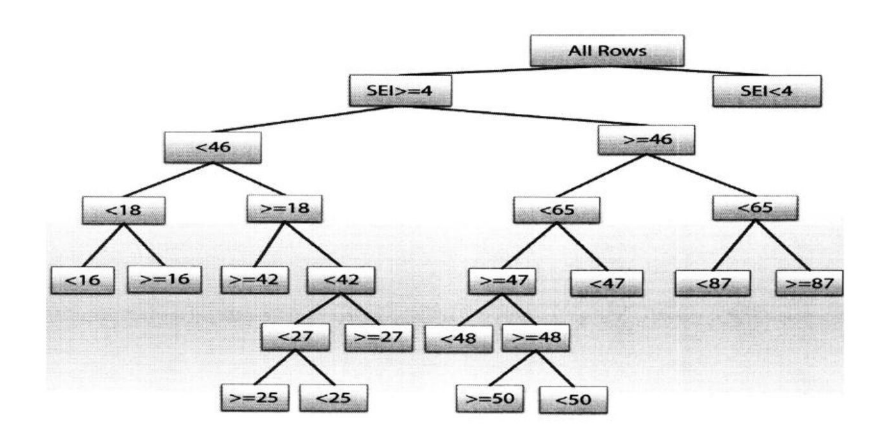
إن عملية الانفصال (Discretizing) بهذه الطريقة، تعزز القوة الإيضاحية (Explanatory Power)، بنسبة 35٪ من $R^2=0.1438$ إلى $R^2=0.1945$، ولكن هل الانفصال مفيد بمجرد كوننا نستخدم فقط متنبئاً وحيداً؟ وهل سيبقى المكسب في القوة التنبؤية قائماً بعد إضافة المتغيرات المشاركة (Covariates)، أو هل ستكون المتغيرات الإضافية ذات الصلة، قادرة على أداء العمل الذي قامت به عملية الانفصال؟

نجيب عن هذا السؤال في الجدول رقم 8.7 من خلال إضافة بعض المتغيرات المشاركة المعيارية الأجور. ونبدأ أولاً بالشكل التربيعي للعمر. وهذه الإضافة، تعزز نسبة التباين الذي تم شرحه بشكل كبير، ولكن يبقى المكسب واضحاً بين النموذج مع وجود قيمة خطية وحيدة بالنسبة إلى المؤشر السوسيو اقتصادي والشكل المنفصل. ويعزز التحصيل التربوي R^2 في نموذج المؤشر السوسيو اقتصادي الخطي بـ 0.037، وفي نموذج المؤشر السوسيو اقتصادي المنفصل بـ 0.030. ويبقى الفرق في R^2، مع النموذج المنفصل الذي نشرح بنسبة تباين تصل إلى 3٪. وفي الأخير، نضيف افتراضات بالنسبة إلى الجنوسة، والعرق، مما يعزز أكثر، القوة الإيضاحية للنموذجين كليهما. وقد تقلص الفرق أكثر في R^2 بين النماذج إلى حوالي 0.018.

وهل يمثل هذا فرقاً كبيراً؟ وهل - حقيقة - عملية الانفصال مهمة جداً إلى هذه الدرجة؟ نؤكد أهميتها لعدد من الأسباب.

أولاً: تبقى الفوارق في القوة التنبؤية حتى بعد إضافة بعض المتنبئات الأكثر قوة للدخل.

ثانياً: قمنا بعملية الانفصال بمتغير واحد فقط، وهذا يحسن - مع ذلك - دقتنا التنبؤية بشكل كبير.



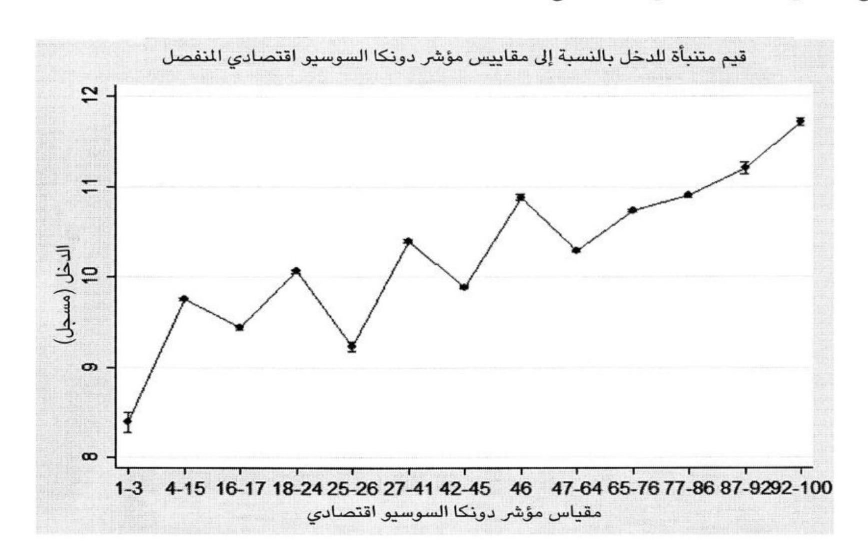
الشكل رقم 4.7: استخدام شجرة التقسيم لتقسيم مؤشر دونكا السوسيو اقتصادي (المكانة المهنية) إلى ثلاثة عشر خانات.

الجدول رقم 7.7: معدل الدخل بفئة المؤشر السوسيو اقتصادي المنفصل.

معدل المؤشر السوسيو اقتصادي	متوسط الدخل $
1–3	11,438.63
4–15	26,338.88
16–17	18,474.6
18–24	31,989.98
25–26	17,198.57
27–41	43,962.59
41–45	29,192.34
46	65,166.95
47–64	42,957.09
65–76	67,200.71
77–86	73,893.76
87–91	90,917.83
92–100	170,696.00

ولكن، من الأهمية أكثر، أن عملية الانفصال قد كشفت عن بعض السمات المثيرة للعلاقة بين الدخل والمكانة المهنية، التي تعتبر مجرد ضجيج إحصائي في

التحليلات المعيارية. ونريد أن نقر النمط العام بمقاييس بديلة للمكانة، وبتنبؤات مأخوذة من العينة، ولكن ربما على الرغم من وجود علاقة خطية إيجابية عامة بين المكانية والدخل، هناك مقايضات محلية صغيرة، حيث تؤدي المهنة المرموقة أجراً أقل - إلى حدّ ما - من مهنة أقل مقاماً.



الشكل رقم 5.7: القيمة المتنبأة للدخل الشخصي

بالنسبة إلى مؤشر دونكا السوسيو اقتصادي المنفصل (Discretized).

الجدول رقم 8.7: إضافة متغيرات منفصلة قد تمَكّن من تحسين التنبؤ (مقاسة بـ R^2).

	المؤشر السوسيو اقتصادي فقط		إضافة العمر		التحصيل التربوي		إضافة الجنوسة والعرق	
	مؤشر دونكا السوسيو اقتصادي مستمر	مؤشر دونكا السوسيو اقتصادي منفصل	مؤشر دونكا السوسيو اقتصادي مستمر	مؤشر دونكا السوسيو اقتصادي منفصل	مؤشر دونكا السوسيو اقتصادي مستمر	مؤشر دونكا السوسيو اقتصادي منفصل	مؤشر دونكا السوسيو اقتصادي مستمر	مؤشر دونكا السوسيو اقتصادي منفصل
R^2	.1438	.1945	.2544	.2916	.2912	.3211	.3371	.3559
R^2 معدلة	.1438	.1945	.2544	.2915	.2911	.3210	.3370	.3558
RMSE	0.9987	0.9687	0.9320	0.9085	0.9088	0.8894	0.8788	0.8663

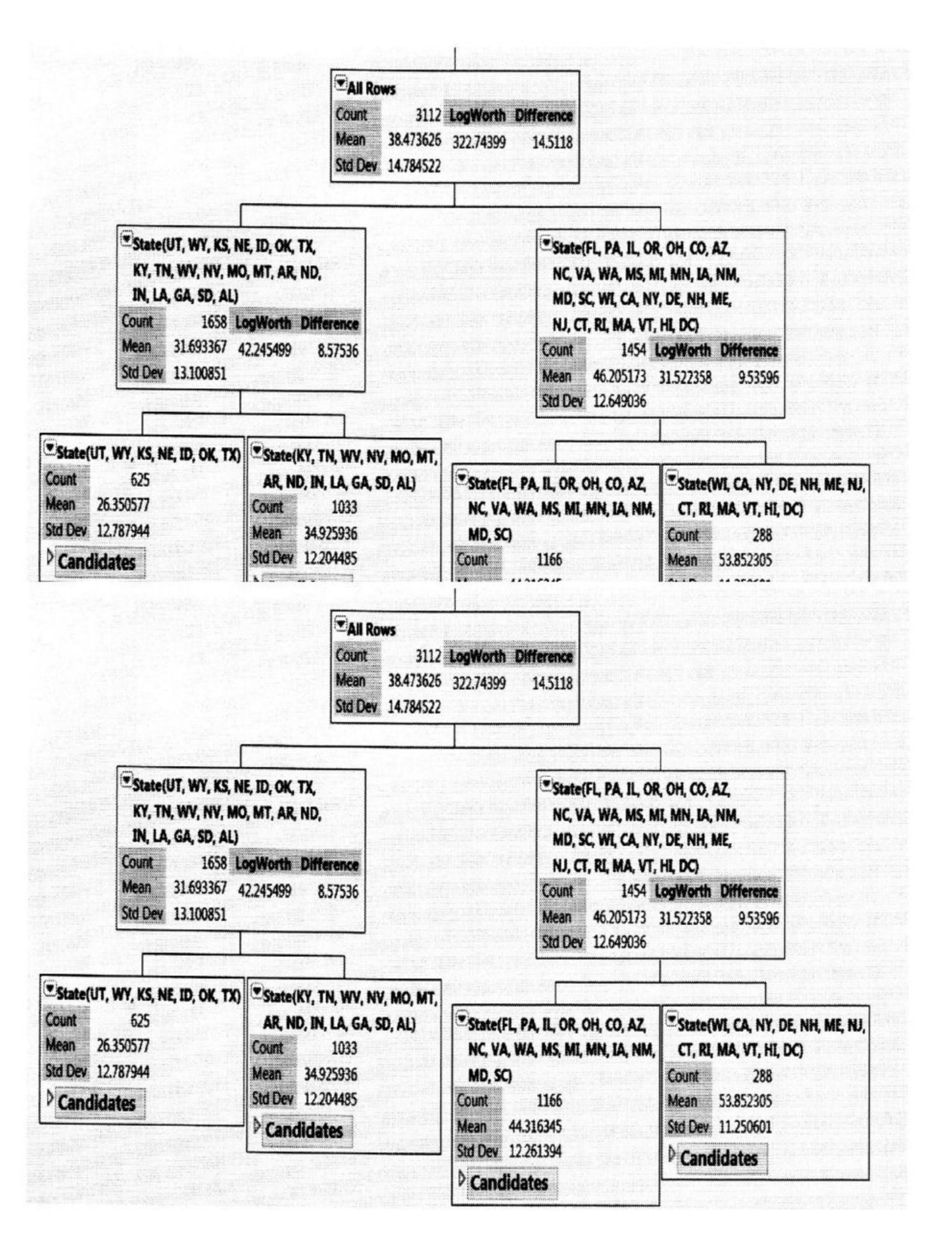
RMSE: جذر متوسط مربع الانحراف.

عملية توزيع متنبئات مصنفة

لقد رأينا سابقاً كيف أن أشجار التصنيف يمكن استخدامها في توزيع متغير مستمر بشكل مثالي في علاقتها نتيجة من النتائج التي نحاول تنبؤها. وتعمل هذه الطريقة بشكل جيد في الحالة التي تتميز فيها العلاقة بين المتنبئ والنتيجة باللا خطية المعقدة. كما يمكن أيضاً استخدام الأشجار على نحو مماثل لإنتاج خانات لمتنبئات تصنيفية. لنقل إن متغيراً اسمياً يملك - نسبياً - عدداً كبيراً من الفئات (مثل متغير بالنسبة إلى مهنة ما) التي نود انصهارها في عدد أصغر للغاية، كي نكون مقترين (Parsimoniousness). إننا نود تصنيف المهن - وهذا أمر مثالي بالنسبة إلينا - إلى فئات أقل، على نحو نسعى فيه إلى تحسين قدرة نموذجنا التنبؤي - أو على الأقل، على نحو نستفيد فيه من التقتير أكثر مما نخسر في التنبؤ الأولي. تقليدياً، نجد سبباً نظرياً من وراء ضرورة ضمّ بعض المهن معاً. وبعد ذلك، نستطيع أن نرى إلى أي درجة يكون لهذا التصنيف معنى ضمن نموذج انحدار ما. وإذا لم يعمل مخطط تصنيف معين بشكل جيد، نتخلى - ببساطة - عنه، ونجرب آخر الذي نظنه «ذا معنى».

ويقترح التنقيب في البيانات إمكانية أخرى. ماذا لو أنتجنا تصنيفنا بطريقة تقوم على البيانات بشكل بحت، أي بطريقة تعظم فيها الفئات قدرتنا على شرح التباين في المتغير المستقل أو متغير النتيجة، معاقبين بذلك النموذج على ضمه مَعْلمات إضافية؟ وننتج بعدها توزيعاً للفئات مثالية من حيث مقايضة التقتير - التنبؤ.

ولبيان قصدنا، نستند - من جديد - إلى بيانات انتخاباتنا على مستوى المحافظة لعام 2012. إن وحدات ترصدنا هي المحافظات، بحيث تتمثل كُلّ محافظة في ولاية معينة. لنقل إننا نريد فحص تأثير الولايات في تنبؤ حصة أوباما (Obama) من التصويت حسب الولاية. نستطيع ضمّ 50 ولاية برمتها باعتبارها متغيرات وهمية في نموذج انحدارنا، غير أن هذا غير أنيق. وهناك خيار آخر تم استعماله بشكل عام، ويتمثل في جمع الولايات ضمن فئات أكبر على مستوى الجهة (مثل إقليم ذي تعداد من فئة -4 أو من فئة -9)، أو على مستوى ميزة أخرى (جنوب/ لا-جنوب، الحق-في - العمل، مقابل، الحق - في - عدم - العمل، وهكذا). وقد تصلح تلك الاستراتيجيات لغاياتنا، ولكنها طرق غير مباشرة تماماً، في بلوغ ما نريد في حقيقة الأمر: ولايات يكون فيها معدل التصويت لأوباما على مستوى المحافظة أعلى أو أقل، مجزأة إلى فئات مثالية.



الشكل رقم 6.7: استخدام أجزاء من الشجرة لإيضاح المتغير الاسمي الفئوي (حالة) في الغامب برو.

وننتج شجرة تقسيم في «الغامب برو»، مرة أخرى من خلال اختيار تحليل تقسيم النمذجة (Analyze Modeling Partition). وكل ما نقوم به – بعد ذلك – اختيار

متغيرنا التابع (حصة أوباما من التصويت)، ومتنبئ مستقل واحد (ولاية). وبما أن متغير الولاية يتميز في «الغامب برو» بكونه فئوي، فإن شجرة التقسيم ستختبر كُلّ الطرق الممكنة لمزج الولايات داخل مجموعتين، واختيار الولاية التي تنتج المجموعتين المختلفتين قدر الإمكان من حيث قيمة متوسط المتغير التابع (حصة أوباما من التصويت).

وإن القسم الأول، المبين في الشكل رقم 6.7 ينتج مجموعة واحدة من الولايات التي تملك محافظاتها متوسط حصة أوباما من التصويت بنسبة 31.7٪، وتملك مجموعة أخرى من الولايات 46.2٪. وتضم المجموعة الأولى، الولايات الأكثر احمراراً من كُلّ الولايات الحمراء - أوكلاهوما، وتكساس، وآوتاه، وأركانساس - في حين تضم المجموعة الثانية كُلّ الولايات الزرقاء إضافة إلى عدد من الولايات الحمراء بشكل قوي مثل جنوب كارولينا، والميسيسيبي. وتقسم مجموعة ثانية من التقسيمات الولايات إلى تقسيمات فرعية تصل إلى أربعة مجموعات بمتوسط حصص التصويت على مستوى المحافظات بلغ 27٪، و35٪، و44٪، و53٪ على التوالي. وقد فصل التقسيم الثاني في الجهة «اليمنى»، الولايات إلى تلك التي تعد ديمقراطية بشكل موثوق فيه (مثل جزيرة رود، ونيويورك، وكاليفورنيا، وهاواي) وولايات ذات ساحة معركة أكثر سخونة (فلوريدا، بنسلفانيا، ونيو مكسيكو). ولاحظ أنه على الرغم من أن هذا التميز لا يخبرنا بالقصة كاملة - ولاية ألينوي الزرقاء بقوة، مثلاً، في هذه المجموعة التي تشكل ساحة معركة، كما هو الحال بالنسبة إلى ميسيسيبي، معقل الجمهوريين. أما التقسيم في الجهة «اليسرى»، فقد فصل الولايات إلى حمراء جداً (آوتاه، وكانساس، وإيداهو)، وغير حمراء بشكل قوي (جورجيا، وكنتاكي، وإنديانا).

الجدول رقم 9.7: توزيع الولايات.

مجموعة	الولايات	متوسط حصة أوباما من التصويت
1	HI, DC	76.39٪
2	RI, MA, VT	63.52٪
3	DE, NE, NH, ME, NJ, CT	55.12٪
4	WI, CA, NY	51.22٪

46.08٪	AZ, NC, VA, WA, MS, MI, MN, IA, NM, MD, SC	5
41.22٪	FL, PA, IL, OR, OH, CO	6
36.96٪	AR, ND, IN, LA, GA, SD, AL	7
32.36٪	KY, TN, WV, NV, MO, MT	8
27.68٪	ID, OK, TX	9
27.68٪	WY, KS, NE	10
18.35٪	UT	11

الجدول رقم 10.7: تأثير توزيع المتغيرات المستمرة إلى خانات في R^2.

جذر متوسط مربع الانحراف	R^2 معدلة	R^2	النموذج
12.01	.3392	.3496	افتراضات الولايات
11.98	.3427	.3448	فئات التنقيب في البيانات
13.53	.16.15	.1637	منطقة التعداد
9.94	.5468	.5487	ضوابط إضافية فقط
8.23	.6895	.6918	ضوابط إضافية + فئات التنقيب في البيانات
8.66	.6563	.6586	ضوابط إضافية + منطقة التعداد

ويسمح «الغامب برو» ببناء شجرة لتعظيم المواءمة في عينة الصلاحية المتبادلة، ولكننا نسعى هنا إلى القيام بشيء مختلف قليلاً. إننا نحاول تعظيم التنبؤ والتقتير في آن واحدٍ عوض منع الإفراط في التدريب. وفي «الغامب برو»، يتم ذلك من خلال بناء شجرة شيئاً فشيئاً، مع فحص إحصاء تطابقي/ تناسبي عقب كُلّ تقسيم. ونفحص حركة معيار أكايكي للمعلومة (Aikake Information Criterion)، التي تقيس

التناسب، وتعاقب نموذجاً ما لإضافته المَعْلمات. ولأن قيم معيار أكايكي للمعلومة يشير إلى تناسب أفضل، فإننا نبني الشجرة ما دام معيار أكايكي للمعلومة يستمر في الهبوط. وعندما يبدأ في الصعود مرة أخرى، نشذب الشجرة ثانية إلى نقطة كان فيها معيار أكايكي للمعلومة الأدنى.

وإن القيام بذلك يولد 11 فئة من الولايات، كُلّ بمتوسط قيمة مختلف بالنسبة إلى حصة أوباما في التصويت. وقد تم تمثيل هذه الفئات الإحدى عشر في الجدول رقم 9.7، من الحصة الأكبر إلى الأصغر من حصص متوسط التصويت لدى أوباما على مستوى المحافظة.

ومن الواضح وجود بعض الولايات النشاز في الجانبين معاً (هاواي، وواشنطن د. س.، في الجانب الموالي لأوباما، وأهوتاه في الجانب المعارض له)، وسينتهي الحال إلى وجود مجموعات صغيرة جداً، ومجموعات كبيرة في الوسط. ولا بُدَّ من الإشارة إلى أن الولايات تتجمع إقليمياً إلى حدّ ما. وكل ولايات بريطانيا الجديدة - في النهاية - توجد في المجموعتين الثانية والثالثة، في حين تظهر ولايات جنوب المحيط الأطلسي في المجموعتين الخامسة والسابعة.

ويقارن الجدول رقم 10.7 الدقة التنبؤية لتجمّع الولايات الذي توصلنا إليه عبر التنقيب في البيانات مع ذلك الذي حصلنا عليه باستخدام تصنيف مستلم مثل منطقة التعداد. إن لدى تصنيفنا 11 فئة، وهناك 9 مناطق تعداد فقط، ومن ثم، فمن المفيد التركيز على قياس تناسب نموذج مثل R^2 المعدلة (الذي يعاقب نموذجاً ما على ضمه معلومات إضافية) من أجل مقارنة عادلة. كما نقارن أيضاً تصنيفنا، مقابل نموذج يضم افتراض واحد بالنسبة إلى كُلّ ولاية (ومن ثم، واحد ذو 50 فئة). إن تصنيف التنقيب في البيانات يتفوق بشكل واسع على مخطط تصنيف منطقة التعداد، مفسراً مرتين التباين في حصة التصويت بشكل عام، ولكن يبقى جذر مربعه R^2 المعدل أقل انخفاضاً قليلًا. وبالتالي، من خلال استخدامنا التنقيب في البيانات، نكون قادرين تقريباً على تفسير قدراً من التباين في المتغير التابع مع نموذج أكثر تقتيراً بكثير.

ولمزيد من فحص هذا، نضيف مجموعة أخرى من متغيرات ضابط تنبؤي للغاية[2]، في أسفل الصفوف الثلاثة لجدول رقم 10.7. ونقوم بذلك لاختبار إمكانية معرفنا لمكان المحافظة، لا يوفر أي معلومة، تعجز متغيرات أخرى عن وصفها. وفي الحقيقة، عندما ندير انحداراً ما الذي يتنبأ بحصة أوباما من التصويت ويضم فقط متغيرات الضابط الديموغرافي، نكون قد فسرنا قدراً لا بأس به من التباين – 54٪. وعندما نضيف متغيرات منطقة التعداد، نكون قادرين على تفسير نحو 11٪ أكثر من التباين، وعادة ما يكون ذلك كافٍ. وسنرى في القوة التنبؤية المحسّنة للنموذج الذي يضم منطقة التعداد، ونختم بأهمية المناطق في مغزى آخر. ولكن في هذا المثال، لدينا أيضاً التصنيف «المثالي» للولايات انطلاقاً من التنقيب في البيانات. ويعد استخدام هذا التصنيف أفضل، حتى في نموذج ذو ضوابط كبيرة: ترتفع R^2 المعدلة من 0.65 إلى 0.68.

استخدام أشجار التقسيم لدراسة التفاعلات

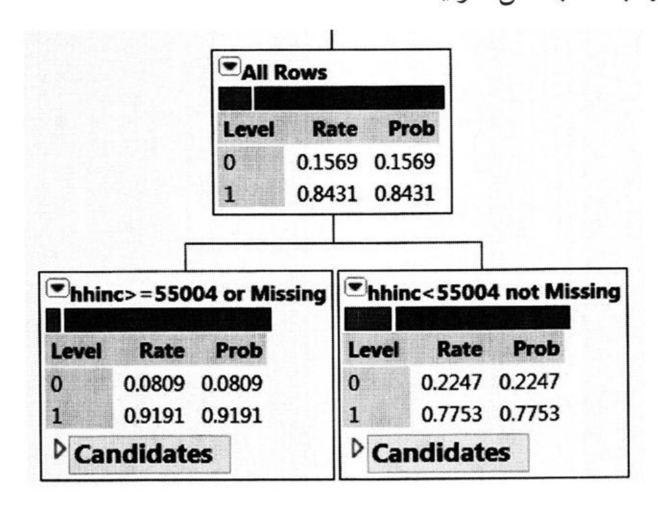
ثمة استخدام آخر لأشجار التقسيم، ويتمثل في تحديد التفاعلات المهمة بين المتغيرات. وتشجعنا نمذجة الانحدار التقليدي على التفكير في العالم باعتباره مكوناً من سلسلة من خصائص إضافية: فاحتمال توظيف شخص ما يشكل دلالة خطية لجنوسته، وعرقه، واعتماده التربوي، وعمره، وتاريخ أعماله السابقة، مثلاً. كما يشكل الدخل دالة إضافية للعمر (تربيع)، والتجربة، والتمدرس.

وأحياناً، نقر بكون المتغيرات تتفاعل لإنتاج النتائج. وربما يتفوق تأثير التعليم في الدخل، على جنوسة شخص ما؛ أو ربما يتوقف تأثير معدل البطالة في احتمال

(2) تتجلى حالات الضبط الإضافية لدينا، في ثلاث قياسات من التوزيع العمري (نسبة الساكنة أقل من 18، ونسب عمرية بين 18-34، ونسبة 65 أو أكبر)، وثلاث قياسات من التوزيع العرقي (نسب البيض غير الهيسبانك، والسود، واللاتينيين)، وثلاث قياسات من التوزيع التربوي (نسبة السكان البالغين الحاصلين على شهادة أقل من شهادة الثانوية، ونسبة من كان يحضر في الكلية، لكنه غادر دون حصوله على شهادة، ونسبة من حصلوا على درجة الباكالوريوس، أو درجة أكبر منها)، ونسبة البطالة بين الذكور، ونسبة الفقر، ونسبة اليد العاملة في المهن الفنية والإدارية، ونسبة الساكنة البروتستانتية الأنجلكانية (المراجع).

إعادة الانتخابات الحالية على ما إن كانت البلاد تعيش حالة حرب أم سلم. فبالنسبة إلى الجزء الأكبر، نتعامل مع التفاعلات على هذا النحو تماماً: إن التفاعلات في اتّجاهين معقدة بقدر ما نسمح للعالم أن يظهر في نماذجنا.

إن أشجار التقسيم يمكننا من البحث عن التفاعلات الأكثر تعقيداً. ولمعرفة السبب، نحتاج إلى الإشارة بشكل مختصر إلى معرفة الشيء الذي تقوم به الأشجار (مزيداً من التفاصيل تجدونه في الفصل 10). إن لوغاريثمات الشجرة تقسم الحالات في بياناتنا، إلى مجموعتين متجانستين قدر الإمكان من حيث النتيجة. وتقوم بذلك من خلال تجريب كُلّ قيمة ممكنة لكُلّ متغير مستقل، وإيجاد السبيل الأفضل لتقسيم العينة إلى مجموعات فرعية. وبعد قيامها بالتقسيم الأول، تكرر العملية مرات عديدة، فتنتج كُلّ مرة تجميعات متجانسة بشكل متزايد على مستوى النتيجة أو المتغير التابع، ومختلفة عن بعضها بعضاً بشكل متزايد.



الشكل رقم 7.7: شجرة حالة التأمين الصحي - التقسيم الأول للبيانات.

وبما أن المتغيرات المتنوعة تم انتقاؤها لإنتاج الحالات التي انتهت في الأخير بتجميعها في «أوراق» نهائية، يمكننا التفكير في كُلّ ورقة محددة بتفاعل معقد من

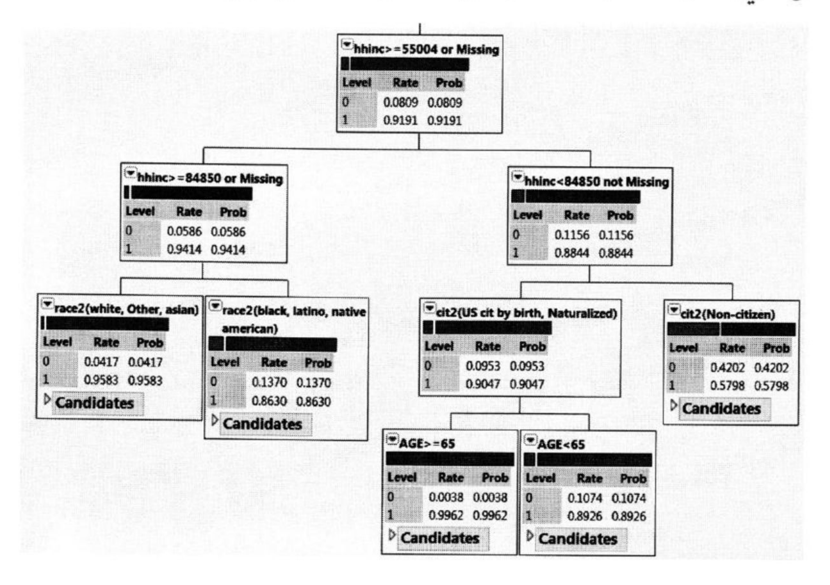
القيم. وبهذه الطريقة، تساعدنا الأشجار على استكشاف تفاعلات معقدة يصعب علينا تحديدها.

ونبين هذا باستعمال بيانات مسح المجتمع الأميركي لعام 2010. ونترك شجرة ما تنمو مستعملين كمتغير تابع، مؤشر الحصول على تغطية التأمين الصحي، والعمر، ومستعملين العرق، والمواطنة، والجنوسة، ودخل الأسرة، والعمر، والتحصيل التربوي، والحالة الوظيفية، والحالة الاجتماعية، باعتبارها متغيرات مستقلة. ونستعمل تقسيم في وقت واحد للحصول على شجرة متواضعة من حيث الحجم. وبخلاف ما كان في القسم السابق، لم نوازن البيانات حول النتيجة، وطبقنا وزن السكّان. وإن الاحتمالات المشروطة الناتجة، تشكل - إذن - تقديرات كميات السكّان.

وفي الشكل رقم 7.7، نرى أن في عقدة الجذر (كل الحالات)، لدى حوالي 83.3٪ من أصل كُلّ الحالات، شكلاً من أشكال التأمين الصحي، وحوالي 15.7٪ لا تملك ذلك. وإن تلك الحالات الموجودة في الأسر ذوي الدخل المختلط الذي يصل إلى 55,000$ على الأقل كُلّ عام، تبلغ 92٪ من المؤمّنين؛ وفي الأسر ذوي الدخل الأقل انخفاضاً، بلغ معدل التأمين حوالي 78٪ تقريباً.

ويمكننا متابعة هذين الفرعين أكثر (الشكل رقم 8.7). وبين المجموعة الأكثر ثراء، تستمر الشجرة في التمييز حسب حالة التأمين الصحي. وتم تقسيم ثانٍ في الدخل في حدود حوالي 85,000$. وبين المجموعة ذات الدخل الأقل انخفاضاً نسبياً (الدخل بين حوالي 55,000$ و85,000$)، تعد المواطنة، الأكثر تنبؤاً للتأمين. وحوالي النصف من غير المواطنين في هذه المجموعة ذات الدخل المرتفع، تفتقر إلى الوصول على التأمين الصحي حسب هذه البيانات. وهذا مختلف جداً عن مواطني الولايات المتحدة في المجموعة ذات الدخل نفسه، إذ يملك حوالي ما يزيد من 90٪، تغطية صحية. ويبرز ذلك، تفاعلاً بين الدخل والمواطنة داخل الأسر ذات الدخل المتوسط العالي. وبين المواطنين على مستوى هذا الدخل، هناك تقسيم في سنّ 65: حوالي 100٪ من الفئة الأكبر سناً مؤمّنة. ومن بين أولئك الذين يتجاوز دخلهم 85,000$، من بيض، وآسيويين وأفراد من «أعراق أخرى»، هناك حوالي 95٪ من المؤمّنين، في حين لا تتجاوز نسبة التأمين لدى السود، واللاتينيين، والأميركيين الأصليين 86٪.

وإذا ما اتبعنا الفرع الأيمن (غير المعروض)، بما في ذلك أولئك الذين لهم دخل أقل من 55,000 $، فسنجد أن بداية أي عمر هو المتنبئ الأهم في حالة التأمين. أما أولئك المؤهلين للرعاية الصحية، وتتجاوز أعمارهم 65+، فهم عملياً مؤمنين بشكل عام. وأما غير كبار السن، فيقسمون بعدها مرة أخرى عند سنّ الرشد. ولا يستفيد من التأمين إلا حوالي 12٪ من أطفال الأسر ذوي الدخل أقل من 55,000 $ سنوياً، ولكن يستفيد حوالي ثلث البالغين في سنّ العمل. ومع ذلك، إن المواطنة بين الأطفال، تتنبأ بالتأمين بشكل كبير: ولا يفتقر إلى التأمين من المواطنين الأميركيين في حدود هذا الدخل، إلا حوالي 10٪، بينما يستفيد حوالي النصف من القاصرين غير المواطنين من هذا التأمين. ومن بين الراشدين في سنّ العمل، نجد تقسيمات من جديد في الدخل (في حدود 35,000 $)، والمواطنة، والحالة الوظيفية.



الشكل رقم 8.7: تقسيم آخر لبيانات موجودة بين هؤلاء ذوي الدخل الذي يزيد على 55,000$.

وعموماً، ننتج شجرة ذات 13 ورقة. وتضم هذه الأوراق نسباً مختلفة من العينة، ولها نتائج مختلفة بشكل كبير من حيث التأمين الصحي كما هو موصوف في (الجدول رقم 11.7).

ومن المهم الإشارة إلى أن أشجار التقسيم، لا تنتقي فقط المتغيرات التي تميز بشكل أفضل بين من هو مؤمّن، ومن هو دون ذلك. وإنها تقوم بذلك بالطريقة التي تنظم بها بشكل جيد، كُلّ الحالات في هذه الخانات. وهكذا، تأخذ بعين الاعتبار نسبة الحالات التي تقع في المجموعات المحددة بالأوراق. وإن القيام بتقسيم يقسّم البيانات - بشكل واضح - إلى فئات نتيجة، ولكن يقسّم حصة صغيرة جداً من البيانات إلى مجموعة واحدة، يكون أقل احتمالاً من تقسيم يولّد مجموعتين ذات متوسط قيم أكثر مماثلة للنتيجة، ولكنها تنتج مجموعتين كبيرتين نسبياً. وهناك ملاحظة أخرى جانبية، تتلخص في إمكانية أن تكوّن المجموعات التي لها معدلات تأمين منخفض، صغيرة جداً. ومن ثم، تكون غالبية غير المؤمّنين ضمن مجموعات ذات مستويات تأمين معتدلة. مثلاً، إن المواطنين الموظفين في سنّ العمل ذوي الدخل المنخفض غير مؤمنين في حدود نسبة تتراوح 32٪، ويفسرون نسبة 27.5٪ من كُلّ أولئك الذين يفتقرون إلى التأمين.

وهل يمكننا استخدام هذه النتائج لتحسين النمذجة التنبؤية؟ ففي الجدول رقم 12.7، نبين عدداً من نماذج الانحدار اللوجيستي الذي يتنبأ تغطية التأمين الصحي. وتذكّر أن في هذه البيانات، تتوفر أكثر من 80٪ من الحالات، على التغطية الصحية، ليكون بذلك انهيار المتغير التابع مائلاً (Lopsided) جداً. وفي هذا النوع من الحالات - وكما أشرنا إلى ذلك سابقاً - يمكن لنموذج ما أن يظهر نجاعته بشكل كبير في تصنيف الحالات بشكل صحيح، من خلال تنبؤ نتيجة إيجابية بالنسبة إلى كُلّ الحالات؛ أي بإمكان نموذج ما، القيام بعمل جيد بخصوص إحصاء تطابقي، يحدد النسبة التي تصنف بشكل صحيح، من خلال تخصيص كُلّ الحالات لنتائج الأغلبية (المؤمّنين في هذه الحالة).

الجدول رقم 11.7: أوراق الشجرة.

الورقة	دخل الأسرة	العمر	الجنسية الأميركية	العرق	الحالة المهنية	نسبة المؤمّنين	نسبة كُلّ الحالات	نسبة غير المؤمّنين جميعهم
1	$85,000	الجميع	الجميع	أبيض، آسيوي، آخر	الجميع	95.84	23.58	6.25
2	$85,000	الجميع	الجميع	أسود، لاتيني، أميركي أصلي	الجميع	86.28	5.04	4.41

0.05	2.02	99.62	الجميع	الجميع	نعم	65+	$55,000-$85,000	3
10.48	15.31	89.26	الجميع	الجميع	نعم	<65	$55,000-$85,000	4
3.11	1.16	57.99	الجميع	الجميع	لا	جميع	$55,000-$85,000	5
0.52	8.61	99.05	الجميع	الجميع	الجميع	+65	<$55,000	6
8.26	12.63	89.74	الجميع	الجميع	نعم	<19	<$55,000	7
1.60	0.54	53.48	الجميع	الجميع	لا	<19	<$55,000	8
14.81	11.45	79.72	الجميع	الجميع	نعم	64-19	$31,000-$55,000	9
5.13	1.48	45.65	الجميع	الجميع	لا	64-19	$31,000-$55,000	10
27.53	13.57	68.18	نعم	الجميع	نعم	64-19	$31,000	11
7.46	2.08	43.74	لا	الجميع	نعم	64-19	$31,000	12
10.38	2.52	35.40	الجميع	الجميع	لا	64-19	$31,000	13

وبالتالي، سيكون لزاماً على نموذج جيد القيام بأفضل من ذلك - سيكون أكثر دقة، ويقوم بعمل لائق للتمييز بين الإيجابيات الصادقة والكاذبة. وسنراقب عدداً من قياسات تناسب النموذج. وفي الحقيقة، إن R^2 - الزائفة (Pseudo-R^2) مقياساً لا بأس به في ضبط مدى أفضلية نموذج ما على تخمين عشوائي في هذه الحالة، لتحديده مدى أفضلية نموذج ما على نموذج صفري (Null Model). إن إحصاءات معيار أكايكي للمعلومة ومعيار بايز للمعلومة ضمّت في التحليل، الرغبة في كُلّ من التنبؤ والتقتير. وكما قلنا أعلاه، إن استخدام النسبة بشكل صحيح، المصنفة باعتبارها معيارنا لنموذج جيد، هو أمر مضلل في هذه الحالة، بسبب انعدام توازن البيانات من حيث النتائج. ونود هنا - في الحقيقة - فحص خصوصية النموذج: نسبة أولئك الذين يفتقرون إلى تأمين، ومصنفين بشكل صحيح على هذا الأساس. والمنطقة في ظل منحنى خاصية التشغيل المتلقي (Receiver Operating Characteristic)، تقيس أيضاً مدى أفضلية نموذج ما على تخمين عشوائي.

الجدول رقم 12.7: تأثير إضافة افتراضات الورقة إلى نموذج تنبؤي.

	نموذج التراجع	R^2 - الزائفة	إحصاء معيار أكايكي للمعلومة	إحصاء معيار بايز للمعلومة	٪مصنفة بشكل صحيح	الحساسية	الخصوصية	المنطقة في ظل منحنى خاصية التشغيل المتلقي
1	متغيرات المكوّن 1	.1492	204,421.9	204,538.8	٪87.30	٪98.40	٪15.13	0.7745
2	افتراضات الورقة فقط	.1749	196,559.1	196,686.7	٪87.06	٪97.65	٪18.25	0.7957
3	متغيرات المكوّن + الأوراق	.2030	189,596.5	189,777.2	٪87.29	98.40٪	15.10٪	0.8182
4	متغيرات ضبط إضافية فقط 2	0.0546	227,415.8	227,596.6	٪86.66	٪100.00	٪0.00	0.6675
5	حالات الضبط + متغيرات المكوّن	.1791	197,291.8	197,578.9	٪87.48	٪98.29	٪17.27	0.7943
6	حالات الضبط + أوراق	.2111	189,788.6	190,086.3	٪87.38	٪98.29	٪16.45	0.8177
7	حالات الضبط + مكونات + أوراق	.2395	182,801.2	183,205.2	٪87.81	٪98.03	٪21.45	0.8342

1. العمر، دخل الأسرة، العرق، المواطنة، الحالة الوظيفية.

2. الجنوسة، والتحصيل التربوي، والحضور المدرسي، والدين، والعالة العائلية.

المصدر: مسح المجتمع الأميركي لعام 2010.

أولاً: ندير نموذجاً يضمّ نسخاً «ساذجة» للمتغيرات المستعملة من لدن شجرة التقسيم أعلاه؛ أي نضمّ فقط العمر (باعتباره متغيراً مستمراً)، ودخل الأسرة (وهو أيضاً متغير مستمر)، والعرق (خمس مجموعات منفصلة)، والحالة المدنية (ثلاث مجموعات: مواطن بالولادة، مواطن مجنّس، ومواطن فاقد للمواطنة)، والحالة المهنية. ويعمل هذا النموذج - في الواقع - بشكل غير سيء تماماً، بل إن المنطقة في ظل منحنى خاصية التشغيل المتلقي، تقترح أفضلية عمله على التخمين العشوائي بنسبة 55٪، وتصنف نسبة 15٪ من غير المؤمّنين، بشكل صحيح. ويعد هذا - جزئياً - شاهداً على العون المقدم من لدن شجرة التقسيم، مع اعتبار أن الشجرة انتقت المتغيرات التي استخدمناها هنا، مشيرة إلى احتمال أهميتها كثيراً. ونقارن هذا بنموذج متغيرات وهمية لكُلّ 13 ورقة، أنتجتها شجرة تقسيمنا (في الحقيقة، 12

ورقة، إذا ما استخدمنا المجموعة الأكبر بمثابة مرجع). ولاحظ أن هذه ليست متغيرات وهمية كما تعودنا على التفكير فيها. إنها تعرّف بتقاطع خمس خصائص، بحيث يقاس اثنان منها بشكل مستمر (ولكن تقسم إلى مجموعات)، أما ما تبقى، فهي قياسات عامة. علاوة على ذلك، لا نستخدم كُلّ التركيبات الممكنة لهذه المتغيرات في التحليل، ولكن نستعمل فقط مجموعات خاصة محددة أعلاه؛ فالمجموعة 1 - مثلاً - تحدد على أساس الدخل والعرق، وتضمّ البيض والآسيويين جميعهم، إضافة إلى أفراد من «أعراق أخَر» ضمن أسر ذات دخل يفوق 85,0000$ سنوياً، بغض النظر عن المواطنة، والعمر، والحالة الوظيفية. ولكن تشترك المواطنة، والعمر، والحالة العائلية في الفوارق بين مجموعات أخرى. وتتميز المجموعة 1 بخاصة عن المجموعة 2 حسب تصنيفات العرق، وعن كُلّ المجموعات الأخرى بنقطة فاصلة في الدخل.

إن النموذج الذي يحتوي فقط على افتراضات «الورقة» هذه، يؤدي، إلى حدّ ما، وظيفة أفضل من متغيرات «المكوّن» - أي التأثيرات الرئيسة غير الخاضعة للتحويل. وعندما نقوم بدمج مجموعتي المتغير هذين، نقوم بشيء أفضل (باستثناء مستوى الخصوصية). ومع ذلك، ليست الفوارق في الدقة التنبؤية كبيرة.

وبعد ذلك، نختبر نموذجاً يحتوي فقط على متغيرات إضافية (الجنوسة، والتحصيل التربوي، والديانة، والحالة الاجتماعية)، لم يتم انتقاؤها بواسطة شجرة التقسيم. ونريد اختبار ما إن كان الامتياز التنبؤي الذي منحته متغيرات الورقة، شيئاً يمكن أداؤه فقط من خلال ضمّ مزيد من المتغيرات المستقلة في النموذج، متغيرات ربما تكون مترابطة بشكل معتدل مع متغيرات المكوّن. وهذه المتغيرات - في حدّ ذاتها - ذات قيمة تنبؤية، وإن كانت بنسبة محدودة (لتمنح R^2 - زائفة تقدر بـ 0.05). وتعرض نتائج هذا النموذج لتأسيس خط أساس جديد. ويمكن أن نرى من خلال فحص الحساسية والخصوصية، قيام النموذج اللوجيستي هنا - في غياب معلومة أفضل - فقط بتصنيف كُلّ الحالات باعتبارها تنتمي لطبقة النتيجة المهيمنة.

وفي النموذجين 5-7، نضيف المتغيرات المستخدمة في النموذجين 1-3 أعلاه. وعندما تضاف متغيرات الضبط إلى متغيرات المكوّن، تقودنا تقريباً إلى الدقة التنبؤية

نفسها التي كانت لدينا لما استخدمنا فقط افتراضات الورقة. وهل هذا يعني أن الأوراق ليست أفضل تماماً من إضافة حالات الضبط؟ يخبرنا النموذج 6، ربما، بعدم صحة ذلك؛ فالأوراق تساهم بشكل كبير في تنبؤ النتيجة على قمة حالات لضبط، وتعمل عملاً أفضل من متغيرات المكوّن نفسها (النموذج 5) من حيث قياسات التناسب.

وأخيراً، نقدم نموذجاً تضم فيه جميع حالات الضبط، والأوراق، والمكوّنات. ولهذا النموذج أكبر دقة تنبؤية من حيث كُلّ قياسات التناسب باستثناء الحساسية التي تعد الأعلى في النموذج 4، ببساطة لأنها خصصت كُلّ الحالات للنتيجة الإيجابية (ومن ثم ضبط 100٪ من الإيجابيات الصادقة). وليست الدقة التنبؤية أكبر بشكل كبير، ولكن قدرتنا على الحصول على امتياز باستخدام قيم تفاعل مولّدة من شجرة التقسيم هو أمر مهم، خاصة إذا ما اعتبرنا أن الطبيعة المائلة للنتيجة، تشكل تحديات أمام أي نموذج تصنيف كان. بالإضافة إلى ذلك، يجب الإشارة إلى أن شجرة التصنيف كان يسمح لها بالانقسام 12 مرة فقط في هذه البيانات. وإذا ما تركنا الشجرة تنمو بشكل كامل، فستنقسم إلى مجموعات أصغر. وقد تكون القيم التنبؤية لقيم التفاعل تحسنت نوعاً ما، إذا ما واصلنا التقسيم.

ويتجلى ضعف طريقة الشجرة - مرة أخرى - في تأويل النموذج النهائي. وعندما تستخدم أشجار التقسيم لبناء قيم التفاعل مثل تلك القيم في هذا التحليل، فإن مُخرجَ نموذج الانحدار لا يمكن قراءته بالطريقة البسيطة نفسها باعتباره نموذجاً تقليدياً. وبعبارات بسيطة، إن «المتغيرات» التي تم تأويلها بشكل عام، باعتبارها قياسات لقوى بارزة اجتماعياً، لم تعد تقوم «بالتمثيل» (Abbott 2001). ولا يمكننا القول «بارتباط 10٪ في الدخل بـ2٪ من الارتفاع في احتمال الحصول على التأمين». نستطيع القول - عوضاً عن ذلك - بربط عضوية في مجموعة ما، المحددة بتوحيد خاص للخصائص، بارتفاع في احتمال الحصول على التأمين؛ أي أننا نسمح للبنية الاجتماعية بجمع الناس بطرق معقدة ضمن مجموعات تشهد نتائج متباية.

ثانياً: في النموذج 7، حيث ضمّنا متغيرات ورقة، إلى جانب مكونات تولّدت منها الأوراق، فإن تفسير المُعامِلات (Coefficients)، إما على مستوى الأوراق أو

التأثيرات الرئيسة، يطرح تحدّ؛ بل لا يمكننا استخدام الطرق بشكل عام لتفسير قيم التفاعل (انظر مثلاً، Brambor, Clark, and Golder 2006; Jaccard and Turrisi 2003). مثلًا، لندرس تفسير المعامل β_2 في المعادلة التالية التي تميز النموذج 7:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_2 G_2 + \sum_{j=3}^{12} \beta_j G_j + \beta_{13}\,income + \beta_{14}\,black + \beta_{15}\,latino + \beta_{16}\,asian + \beta_{17}\,nativeamerican + \beta_{18}\,otherrace + \gamma X + \delta Z + \varepsilon$$

تمثل β_2 تأثير كون الفرد في مجموعة 2، (والمترابطة بكون الفرد في مجموعة 1) في احتمال الحصول على تأمين[3]. ومع ذلك، تحدد هذه المجموعة بعلاقتها بالمجموعة 1 على مستوى العرق (أسود، ولاتيني، وأميركي الأصل، مقابل أبيض، وآسيوي، وآخر). إنه الفرق العرقي بين أولئك الذين يصل دَخْلهم إلى 85,000 $ على الأقل. ومع ذلك، فإن المجموعات العرقية الأخرى تدخل في التحليل باعتبارها حالات ضبط منفصلة (تم قياسها جميعها في مقابل المجموعة المرجعية، الجنس الأبيض). ويعد الدَخل أيضاً متغيراً منفصلاً، يقاس بشكل مستمر. ومن ثم، فإن التأثير في التأمين الصحي للتباين العرقي بين أصحاب الدخل المرتفع، يعد صافٍ من الدخل والعرق. وكي نكون واضحين، إن إيجاد معنى لهذا أمر معقد، ويعد تباين المجموعة الأبسط، في هذه الحالة الخاصة، بما أن المجموعة 1، والمجموعة 2 تشكلان انقساماً واحداً بعيداً عن بعضهما بعضاً. ومن ناحية أخرى، تشكل المجموعة 1، والمجموعة 10 تقسيمات منفصلة، وتتباين من حيث الدخل، والمواطنة، والعرق/ والإثنية، والمجموعة العمرية.

ويطرح التأويل تحد بسبب ضم كُلّ من افتراضات الورقة ومكوناتها في النموذج نفسه. وإن اختيار واحد أو المجموعة الأخرى من هذه المتغيرات ييسر التأويل بشكل كبير. ومن أجل تعظيم التنبؤ - مع ذلك - تبقى عملية ضمّ كُلّ المعلومات في هذه المتغيرات أمراً مفيداً.

(3) تعد المجموعة 1، المجموعة المرجعية؛ فهي لا تظهر - إذن - في المعادلة. وتعد x القوة الموجهة لمتغيرات «المكوّن» المتبقية، وZ القوة الموجهة لمتغيرات «الضبط» الإضافية. (المترجم)

خلاصة

نحن لا نعيش في ما يسميه أندرو أبوت (Andrew Abbot) «حقيقة خطية عامة»؛ وإنما تولّد النتائج من خلال التفاعل المعقد لعمليات اجتماعية التي تعد المتغيرات - عادة - بالنسبة إليها، مجرد اختزال مريح؛ بل إن طريقتنا المعيارية في ترميز تفاعلاتنا بين المتغيرات ليست كافية لضبط تعقيد كيفية تفاعل الخصائص في العالم. وتتلخص النماذج في نماذج الانحدار، التي تساعدنا بدقة عبر التبسيط التي تفرضه على العالم، مشيرة إلى متوسط العلاقات ذات الأهمية الكبرى.

ومع ذلك، نقوم بشيء أفضل من ذلك، في بعض الأحيان، وقد سبق أن أشرنا - في هذا القسم - إلى مدى قدرة أدوات التنقيب في البيانات مثل أشجار التقسيم، السماح لنا بأداء هذا. ويمكن لأشجار التقسيم بخاصة، أن تكشف عن كيفية تفاعل المتغيرات - ويسمى متغير في شجرة التقسيم بعد هذه الجودة الدقيقة: «مربع كاي» للكشف عن التفاعل التلقائي (CHAID). علاوة على ذلك، إن الرفع من قدرة الكشف عن التفاعل، يمكن أن يساعدنا في تحسين القدرة التنبؤية - ومع ذلك، كثيراً ما تتفوق أشجار التقسيم على الانحدار اللوجيستي في مهام التصنيف. وقد بينا هنا أن استخدام قدر صغير من قوة الكشف عن التفاعل لأشجار التقسيم، يمكن أن تحسن أداء نماذج الانحدار على مستوى التنبؤ.

ويمكن استخدام طرق التنقيب في البيانات في إنتاج تحولات متغير جديد - عملية توزيع مثالي للخانات وإنتاج قيم تفاعل معقد. وفي بعض الأحيان، يمكن النظر إلى هذا باعتباره يجعل مجموعة متغيرنا أو سمتنا أكثر تعقيداً. وبعد ذلك سننتقل إلى مجموعة تقنيات من أجل تقليص تعقيد مجموعة سمتنا، مع الحفاظ - في الوقت نفسه - على البنية العامة للبيانات: طرق استخراج متغير.

الفصل الثامن

استخراج المتغيرات

تحليل المكوّن الرئيسي

عندما تكون لدينا بيانات ذات بعد عالي، أي بيانات واسعة جداً (خصائص أو متنبئات كثيرة)، نريد أحياناً إيجاد طرق لتقليص بُعدِيتها (Dimentionality). وقد سبق لنا مناقشة طرق انتقاء السمة مثل الانحدار التدريجي (Stepwise Regression)، واللاسو (Lasso)، وانحدار معامل تضخم التباين (VIF). وتعد هذه الطرق خيارات - لا محالة - عندما نريد تخفيض أبعاد متغيرات المتنبئ على مستوى علاقتها بنتيجة ما. كما تعد أدوات انتقاء السمة، طرقاً «مراقبة» برمتها، مادام هناك بُعد محدد من البيانات (النتيجة، أو الهدف، أو المتغير التابع) يتمتع بامتياز، وأننا ننتقي متغيرات مهمة بالنسبة إلى كيفية علاقتها بهذا المتغير المتمتع بامتياز.

ولكن لا نملك دائماً متغيراً نهتم به بشكل خاص. وأحياناً، لدينا ببساطة كتلة من البيانات، ونريد من خلالها تمييز أنماط في هذه البيانات. ومن الممكن اختصار جزءٍ لا يستهان به لما يُعد مهماً في مجموعة كبيرة من المتغيرات، والتعبير عنه بلباقة وبساطة بواسطة حفنة خصائص ملخصة. ومن أجل هذا النوع من الحالة بالذات الذي طُورت في إطاره هذه التقنيات العتيقة من تحليل المكوّن الرئيسي، وقريبه الوثيق الصلة به - تحليل العامل (Factor Analysis).

لندرس ثلاث متغيرات في مجموعة بيانات على مستوى انتخابات المحافظات لعام 2012:

- متوسط الدخل.
- ونسبة السكان الذين يمتلكون شهادة جامعية أو شهادة أعلى.
- ونسبة القوة العاملة في الوظائف المهنية، والإدارية.

ولا غرو أننا نجد ترابط هذه التصورات الثلاثة فيما بينها، وبالنتيجة، إن معظم الناس في الوظائف المهنية أو الإدارية، هم خريجو الكلية، كما يميل كُلّ من خريجي الكلية وأولئك الذين يشغلون تلك الوظائف إلى أن تكون لديهم رواتب أعلى من المتوسط. ولفحص ترابطاتهم، نستطيع إنتاج مصفوفة ارتباط (Correlation Matrix) (الجدول رقم 1.8)، ورسم بياني للتشتت (Scatterplot) ثلاثية الأبعاد (الشكل رقم 1.8)، بحيث يُعد هذا الأخير إذن من غامب.

الجدول رقم 1.8: مصفوفة الارتباط.

	متوسط الدخل	٪ التعليم العالي	٪ مهني/ إداري
متوسط الدخل	1	–	–
٪ التعليم العالي	0.690	1	–
٪ مهني/ إداري	0.585	0.788	1

يقدم هذا تأكيداً بخصوص ترابط هذه المقاييس على مستوى المحافظة، بما أنها تعد جميعها طرقاً تشير إلى الثراء النسبي لمحافظة ما. يستطيع المرء الآن تقليص الأبعاد - ببساطة - من خلال استعمال إحدى هذه الخصائص، وافتراض أنها مناسبة في التعبير عن مفهوم الثراء. ولكن يجب الأخذ بعين الاعتبار أن المتغيرات غير مترابطة بشكل كامل. ومن الواضح أنها تعبر عن أشياء مماثلة، ولكن غير متطابقة حول المحافظات. وفي المقابل، نستطيع إنتاج خاصية رابعة، تعبر عن معظم التباين في هذه المتغيرات الثلاث، منجزة - بالضبط - القدر نفسه من التقليص البعدي، ولكن من خلال سحب المعلومات من المقاييس الثلاثة جميعها.

ونقوم بهذا، من خلال إيجاد المكوّن الرئيسي الأول لهذه المتغيرات الثلاث. ولكن، ماذا يعني هذا بالضبط؟ لندرس مصفوفة الارتباط، كما وردت في الجدول

1.8؛ فهي الشكل المقعّد لمصفوفة تباين التغاير (Variance-Covariance Matrix)، التي تصف الترابطات بين المتغيرات بطريقة لا تجد حلاً لتباينات المتغيرات أنفسها.

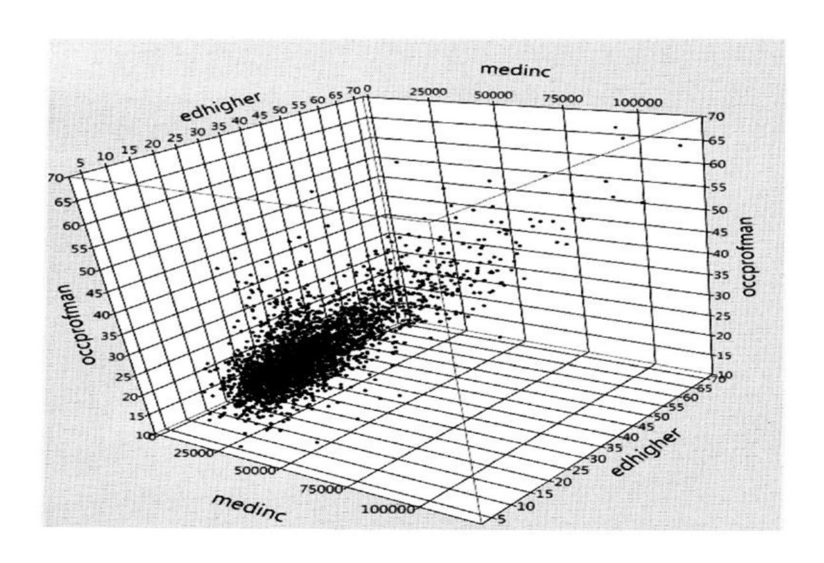
ومن أصل أي مجموعة متغيرات، يمكن للمرء أن يشتق مصفوفة تباين التغاير لبعدية $p \times p$، حيث إن p تمثل عدد المتغيرات قيد الدراسة. ويمكن للمرء أن يحدد لهذه المصفوفة، مجموعة من p متجهات خاصة لـ p تحديداً، تعرف بالمتجهات الذاتية (Eigenvectors)، التي تعيد نفسها مرات عديدة، تعرف بقيمة ذاتية (Eigenvalue)، عندما تُضرب في مصفوفة التغاير. وتمثل هذه المتجهات خطوطاً مستقيمة، تصف التباين على نحو أكثر فاعلية في البيانات عندما يتم إسقاطها عبر سحابة البيانات ذات البعد-p.

لدى كُلّ اتّجاه ذاتي، قيمته الذاتية، تخبرنا الأحجام النسبية بالأهمية النسبية لكُلّ متجهة من المتجهات الذاتية على مستوى وصف تباين البيانات؛ أي إن المتجهة الذاتية ذات القيمة الأكبر، تصف الحصة الأكبر للتباين في البيانات. كما تصف المتجهة الذاتية ذات القيمة الذاتية الموالية الأكبر، الحصة الأكبر للتباين المتبقي بعد ما تمت إزالة التباين الذي وصف من قبل المتجهة الذاتية الأولى؛ وهكذا.

تجدر الإشارة إلى أن هذا يعني أن كُلّ المتجهات الذاتية موجودة في الزاوية القائمة (Right Angle) لبعضها بعضاً؛ مما يعني عدم ارتباطها مع (متعامدة مع Orthogonal to) بعضها بعضاً. وإن ما يطلعوننا عليه، هو أمر مهم للغاية؛ بحيث إذا أخذنا - مثلاً - رسم بياني للتشتت (Scatterplot) الثلاثي الأبعاد في الشكل رقم 1.8، فسنستطيع إدارة سحابة البيانات حول نقطتها الوسطى (Centroid) (النقطة الوسطى)، إذ سيوجد خط من الخطوط التي تقلص المسافات بينها وبين البيانات نفسها على امتداد محور x.

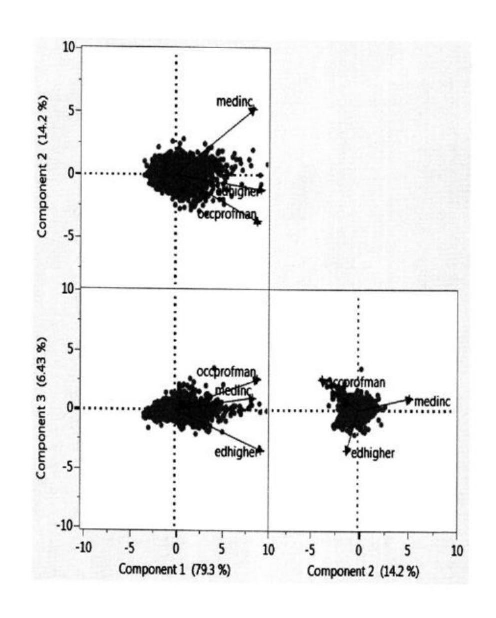
وإن المتجهة التي وصفت هذا الخطّ، ستكون المتجهة الذاتية ذات أكبر قيمة ذاتية، وسيمثل محوري y وz، المتجهتين الأخريين. وبالنسبة إلى هذه السحابة من سحابات البيانات ثلاثية الأبعاد، سيكون محور x المكوّن الرئيسي الأول. أما المحوران y وz، فسيكونان المكوّنين الرئيسيين الثاني والثالث.

وعندما ننجز تحليل المكوّن الرئيسي لمتغيراتنا الثلاثة مستخدمين «الغامب»، يكون بإمكاننا - بصرياً - مراقبة العلاقات بين البيانات، والمتغيرات، والمكوّنات الرئيسة من خلال إنتاج رسوم بيانية ثنائية. وهذه ببساطة رسوم بيانية للتشتت ذات المكوّنات الرئيسة، المشكلة للمحاور (الشكل رقم 2.8).



الشكل رقم 1.8: رسم بياني للتشتت الثلاثي الأبعاد من إنتاج «الغامب برو».

ويعد كُلّ من الرسمين البيانيين الأعلى والأوسط، مكونا 1 - المكوّن الرئيس الأول - باعتباره المحور x. وإن مسألة سحابة البيانات منتشرة أفقياً بشكل واضح، تبين معظم التباين في البيانات على طول هذا البعد - 79.3 ٪ منها، تحديداً (الجدول رقم 2.8). كما نستطيع أيضاً استكشاف أن التباين العمودي، يعد أكثر وضوحاً بعض الشيء في الرسم البياني أعلى اليسار. وهذا راجع إلى كون البعد العمودي هنا، هو المكوّن الرئيس الثاني، الذي يصف التباين - في حدّ ذاته - أكثر مما يقوم به المكوّن الرئيس الثالث.



الشكل رقم 2.8: الرسم البياني الثنائي للمكونات الرئيسة (الغامب برو).

الجدول رقم 2.8: المكوّنات، والقيم الذاتية، والتحميلات.

				التحميل	
	القيمة الذاتية	وصف التباين	متوسط الدخل	٪ التعليم العالي	٪ مهني/ إداري
المكوّن الأول (PC1)	2.73	79.3٪	0.85	0.93	0.89
المكوّن الثاني (PC2)	0.43	14.2٪	0.52	−0.12	−0.37
المكوّن الثالث (PC3)	0.19	6.4٪	0.11	−0.34	0.26

الجدول رقم 3.8: انحدار حصة أوباما من الأصوات حول المكوّنات الرئيسة.

	النموذج 1	النموذج 2	النموذج 3
متوسط الدخل	-0.225 (0.029)	------	------
٪ التعليم العالي	1.148 (0.052)	------	------
٪ الوظيفي/ الإداري	-0.728 (0.061)	------	------
المكوّن الأول PC1	------	1.855 (0.159)	1.855 (0.159)
المكوّن الثاني PC2	------	------	-1.159 (0.375)
المكوّن الثالث PC3	------	------	-11.116 (0.558)
الاعتراض	48.42	38.44	38.44
جذر متوسط المربعات (RMSE)	13.69	28.44	13.69
R^2	.148	.034	.148
Adj. R^2	.147	.036	.147

إن النقاط الرمادية الغامضة في الشكل رقم 2.8، تمثل الحالات الفردية، ويمكننا إدراك أن عملية إدارة سحابة النقطة بحيث يمتد بعدها ذو التباين الأكبر على طول المحور الذي يحدده المكوّن الرئيس الأول. وتبين الأسهم كيفية ارتباط كُلّ من المتغيرات التي تم قياسها، بالمكوّنات الرئيسة، كما يمكن إدراك أن كُلّ المتغيرات مترابطة ارتباطاً وثيقاً بالمكوّن الرئيسي الأول. وإن الأعداد التي تدعى تحميلات المعامل، تطلعنا على مدى ارتباط كُلّ متغير من متغيرات المساهم، بكل مكوّن من المكوّنات الرئيسية. ولاحظ أن كُلّ الارتباطات القائمة بين المتغيرات والمكوّن الرئيسي الأول، أعلى من أي ارتباط من ارتباطات ثنائية المتغير (Bivariate Correlations) بين المتغيرات التي شاهدناها في الجدول رقم 1.8. ويقدم المكوّن

الرئيس الأول، إذن، ملخصاً جيداً لما تشترك فيه هذه القياسات الثلاثة، من دون أن تفضل أي متغير على آخر.

وسيستخدم المكوّن الرئيسي الأول بمثابة طريقة ممتازة في تقليص التعقيد التحليلي، باعتباره قياساً ملخصاً. ويمكن بيان ذلك من خلال العمل على انحدار نسبة التصويت لصالح أوباما، **أولاً** على مستوى المتغيرات المركبة الثلاثة (النموذج 1 في الجدول رقم 3.8)، ثم - ببساطة - على مستوى المركب الرئيسي الأول (النموذج 2). وبعد ذلك، نضيف المكوّنين الرئيسيين (النموذج 3). وتم تقسيم متوسط الدخل على 1,000 بغية تسهيل عملية التأويل.

لاحظ إن لدى المكوّن الرئيسي علاقة إيجابية قوية، على الرغم من أن لدى المتغيرين المستقلين ارتباطات جزئية مع النتيجة (حصة أوباما من الأصوات) السلبية. وهذا يبين جدوى المكوّنات الرئيسة في التخلص من الصعوبات التأويلية التي ولدتها مسألة إدراج المتغيرات المترابطة للمتنبئ في النموذج. ولدى كُلّ المتغيرات على حدة، علاقة إيجابية ثنائية المتغير مع حصة أوباما من الأصوات، ولكن تبقى العلاقة الأقوى في حالة المتغير الذي يقيس نسبة السكّان البالغين الحاصلين على شواهد عليا. ومن ثم، عندما نبقي على نسبة السكّان الحاصلين على شهادة جامعية ثابتة، فإن تأثيرات المتغيرين الأخريين تصبح سلبية. وفي هذا النموذج، نحتاج إلى تحديد - بحذر - مسألة أن العلاقة بين متوسط الدخل وحصة الأصوات مثلاً، هي سلبية فقط بعد ضبط كثافة خريجي الكلية. وإذا ما كنا - مع ذلك - نؤول كُلّ متغير باعتباره عكساً للثراء الرئيسي، فإننا قد نسيء وقد لا نسيء تأويل الدليل على أنه دليلاً ممزوجاً. وإن عملية طيّ المعلومة المشتركة في مكون رئيسي وحيد، يفضي بنا إلى القدرة على بيان علاقة أكثر إيجابية وبساطة بين هذه المظاهر من مظاهر التراتبية الاجتماعية، وحصة أوباما من الأصوات.

وفي الجدول رقم 3.8، نلاحظ أن عملية الانتقال من النموذج 1 إلى النموذج 2 - عندما نتنبأ بأصوات أوباما - تقلص R^2 بنسبة 77 ٪. ولكن كيف يمكن حدوث ذلك إذا كان هذا المكوّن نفسه يصف 78 ٪ من التباين بين المتنبئات الثلاثة؟

أولًا: جواب ذلك أن المكوّن الرئيس لم يُحدَّد استناداً إلى متغير النتيجة، ولكن يصف فقط علاقات بين المتنبئات الثلاثة.

ثانياً: لا ارتباط أي من المتنبئات بشكل كبير مع النتيجة؛ فنسب الارتباطات القائمة بين حصة أوباما من الأصوات هي كالتالي: أما شهادات الكلية، فهي: r = 0.298.، أما متوسط الدخل، فهو r:0.102، وأما الوظائف المهنية أو الإدارية، فهي r = .107. وفي الحقيقة، إن حصة أوباما من الأصوات أكثر ارتباطاً للغاية مع المكوّن الرئيسي الثالث (r = −0.329) من الأول (p=.19290) أو مع أي من المتنبئات بمفردها. وبتعبير مبسّط، إن الثراء مرتبط إيجاباً مع حصة أوباما من الأصوات، ولكن العلاقة ضعيفة. ويساعد تحليل المكوّن الرئيسي – في الحقيقة – على الإفصاح – في هذه الحالة – عن أن غالبية العمل التوضيحي الذي تم إنجازه بواسطة ثلاث متغيرات قيد الدراسة، لم يتم بمعظم ما يشتركون فيه.

وأخيراً، لاحظ أن مقاييس التناسب (جذر متوسط المربعات، R^2، وR^2 المعدلة)، متطابقة في النموذجين 1، و3. وهذا راجع إلى كون المكوّنات الرئيسة الثلاثة جميعها، تضم كُلّ المعلومات في المتغيرات الأصلية المقاسة.

ولبيان خاصية النموذج المبسّط – بشكل حقيقي – تحليل المكوّن الرئيسي، تدعو الحاجة إلى البدء بمزيد من المتغيرات. ونجمع 22 متنبئاً لحصة أوباما من الأصوات، وننجز تحليل مكوّن رئيسي (الجدول رقم 4.8). وإن انحدار حصة الأصوات على مستوى 22 متغيراً، ينتج R^2 من 0.5826 ومع ذلك، فالنموذج معقد جداً، والعديد من المتغيرات مترابطة. ويمكن استعمال تحليل مكوّن أساسي لتقليص بعدية البيانات، مستخدمين هذه المرة، تَحكُّم تحليل المكوّن الرئيسي للستاتا (Stata).

كما يمكننا فحص القيمة المنخفضة للقيم الذاتية من خلال التوسل بالرسم البياني (Scree Plot) بعد التحليل(الشكل رقم 3.8). ونستطيع رؤية انحدار العدد بشكل سريع في البداية، ويستوي عند حوالي خمسة. ويطلعني هذا على أن المكوّنات الخمسة مجتمعة، تصف معظم (حوالي 68.5 ٪) التباينات في المتغيرات 22. ومع ذلك، ومن أجل أخذ الحيطة، سندرج مكونين آخرين – ليصبح لدينا سبعة مكونات إجمالاً.

إن انحدار حصة أوباما من الأصوات على مستوى هذه المكوّنات السبعة، تنتج R^2 من 0.4338، التي تمثل حوالي 75٪ من التباين الأولي الموضح في النموذج بأكمله، وإن كان ذلك بقلة قليلة من المتغيرات. ولأن كُلّ المكوّنات الرئيسة، هي في مستوى واحد (أي إنها عادة موزعة بمتوسط 0 وانحراف معياري 1)، فسيكون بالإمكان مقارنة معاملات الانحدار بشكل مباشر. وفي الجدول رقم 5.8، يمكن ملاحظة أن لدى المركبات 1، و2، و3، و5، علاقات إيجابية مع حصة أوباما من الأصوات، وأن مركبي 1، و2، هي أقوى المركبات. وترتبط هذه المركبات إيجاباً بالكثافة السكّانية، وبنسبة سكّان محافظة ما سوداء (على التوالي)، كما ترتبط سلباً بمتغيرات من قبيل نسبة كبار السنّ، ونسبة القاصرين في السكّان (على التوالي).

ومن الأشياء المفيدة بشأن تحليل المكوّن الرئيسي، تتمثل في حقيقة أن المكوّنات ذاتها غير مترابطة. وبسبب هذه العمودية، تصبح R^2 لانحدار ما على مستوى كُلّ المكوّنات، مجموع قيم R^2، انطلاقاً من الانحدارات على مستوى كُلّ مركب من المركبات على نحو فردي. وكل مركب، يصف قسم فريد من التباين في متغير النتيجة، على الرغم من أن النتيجة لم تكن مدرجة (في الواقع) داخل تحليل المكوّن الرئيسي نفسه. ويخبرنا الجدول رقم 5.8 بأن المكوّن 2، يصف أكثر من 21 ٪ من التباين في حصة أوباما من الأصوات بمفردها، وأن أجزاء كبيرة من التباين، وصفت أيضاً من قبل المكوّنات 3، و6، و7. وإن العديد من المكوّنات غير مترابطة بشكل أساسي بمتغير النتيجة، وهو أمر متوقع بالنظر إلى أن النتيجة لم تستخدم في توليد المكوّنات.

الجدول رقم 4.8: نتائج تحليل مكون أساسي.

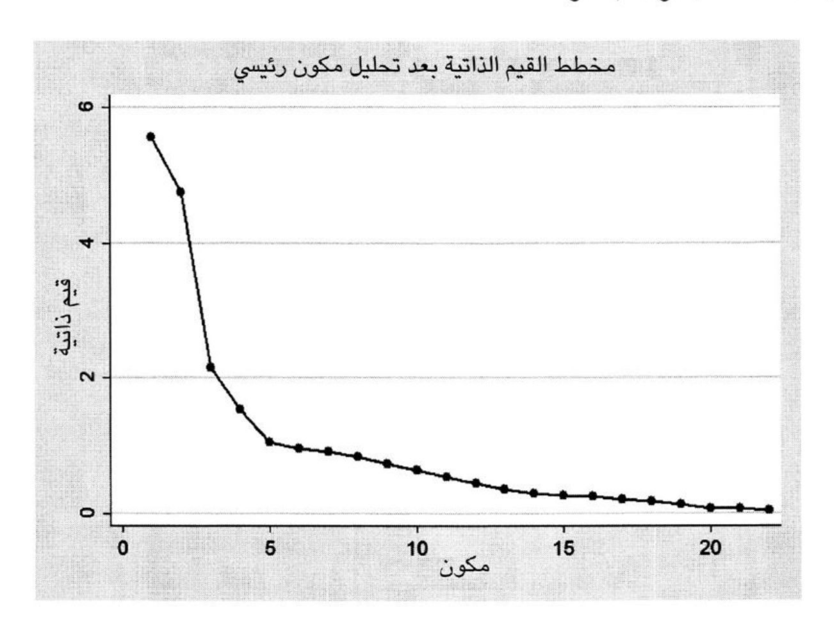
المكوّن	القيمة الذاتية	نسبة التباين	تباين تراكمي
1	5.56	0.253	0.253
2	4.75	0.215	0.469
3	2.15	0.098	0.567
4	1.53	0.070	0.637
5	1.05	0.048	0.685

0.729	0.044	0.97	6
0.770	0.041	0.91	7
0.808	0.037	0.83	8
0.841	0.033	0.73	9
0.871	0.029	0.65	10
0.896	0.025	0.54	11
0.916	0.020	0.44	12
0.932	0.016	0.36	13
0.945	0.014	0.30	14
0.957	0.012	028	15
0.986	0.012	0.26	16
0.978	0.009	0.20	17
0.986	0.007	0.17	18
0.992	0.006	0.14	19
0.995	0.003	0.07	20
0.998	0.003	0.07	21
1.000	0.002	0.04	22

ملاحظة: مدخل التغيرات؛ N = 3.114

ويفضي بنا هذا، إلى العائق الرئيس لتحليل المكوّن الرئيسي: تأويل المكوّنات الفردية ذاتها. وتصف المكوّنات أجزاء فريدة من المعلومات المدرجة في كُلّ المتغيرات - ولكنها غير مضمونة - في سياق متعدد الأبعاد مثل هذا، إلى درجة أن معظم المتغيرات، أو أي من المتغيرات ستكون ذات تحميل عالي على مستوى إحدى المكوّنات. وفي هذه الحالة، يكون تحميل العديد من المتغيرات متوسطاً فقط على مستوى أي من المكوّنات، ولكن تحميلها مماثل على مستوى اثنين أو ثلاثة.

ومن ثم، فإن تأويل «معنى» مكوّن ما عند أداء تحليل مكون رئيسي مع متغيرات عديدة، عادة ما يكون مباشراً.



الشكل رقم 3.8: رسم بياني يوضح قيماً ذاتية لمكونات من تحليل المكوّن الرئيس.

وتتمثل النقطة الرئيسة في أن الاستعمال الرئيسي لتحليل مكوّن رئيسي، هو تقليص في البعدية ذاتها، مُشَكِّلةً نموذجاً أكثر تقتيراً. ويمكن أن يأتي هذا على حساب تأويل يسير. إذا أردنا تقليص البعدية بالتزامن مع تحسين القابلية التأويلية، فستكون إحدى الاستراتيجيات الأفضل هو أداء تحليلات عامل، أو تحليل المكوّن الرئيسي على مستوى المجموعات الفرعية للمتغيرات، كما فعلنا مع تحليل المكوّن الرئيسي لمتوسط الدخل، وللتحصيل العلمي، ونسبة الوظائف المهنية والإدارية التي تصدّرت هذا القسم. كما يؤدي هذا إلى مزيد من المتغيرات الملخصة القابلة للتأويل التي يتم توليدها. ومع ذلك، إذا تم استخدام مجموعات منفصلة لمتغيرات مترابطة ارتباطاً نظرياً لتوليد عوامل منفصلة، فسترتبط - على الأرجح - هذه العوامل ذاتها. وبالنتيجة، لن تصف هذه المجموعات مكونات منفصلة للتباين في المتغير التابع، وأن ارتباطاتها نفسها، ستحتاج إلى الفحص كجزء من تحليل عام.

ملخص

في هذا القسم، ركزنا على الطريقة الأكثر شيوعاً من طرق استخراج المتغير - تحليل المكوّن الرئيسي. وهذه الطريقة ليست طريقة من طرق التنقيب في البيانات في جوهرها، ولكنها تستخدم بشكل اعتيادي من قبل متخصصين في التنقيب في البيانات، لجعل مجموعات البيانات المعقدة أكثر قابلية للطّرق. وإنها لطريقة صارمة ودقيقة لتلخيص غالبية التباين المشترك بين مجموعة كبيرة من المتغيرات ذات عدد أصغر من المقاييس. لقد بينا هنا أنه فقط 7 مكونات رئيسية، هي القادرة على إنجاز حوالي ثلاثة أرباع و22 من المتغيرات المنفصلة التي اشتقت منها في تنبؤ حصة أوباما من الأصوات لعام 2012 على مستوى المحافظة. وهناك بالطبع، مقايضة للدقة من أجل التقتير - التي تعد أكثر أهمية - تتوقف على ما تبحث عنه من نموذجك.

تحليل المكوّن المستقل

إن تحليل المكوّن المستقل (Independant Component Analysis) (ICA) - الذي تمّ في الأصل تصوره من قبل عالم الحاسوب بيار كومون Pierre Comon) (1994 - ينحدر من تحليل المكوّن الرئيسي، الذي يشبهه قليلاً. ويحرك الاختلافات بين التقنيتين، أنواع المشاكل التي تمّ تقديمها لها في البداية بغرض إيجاد حلّ لها، والتي تناسبها بشكل أفضل. وربما تستخدم التقنيتان كلاهما باعتبارهما أداتان من أدوات تقليص البيانات أو التبسيط، لاستكشاف بنيات البيانات الرئيسة ضمن بيانات معقدة متعددة المتغيرات.

وقد تستعمل التقنيتان كلاهما أيضاً، لحل مشكلة غير مختلطة (Unmixing) أي فرز الإشارات المستقلة المختلطة معاً في بيانات الترصد. ويمكن استخدام تحليل المكوّن المستقل في الحالة الثانية، ولكنها تستخدم بشكل أنسب بكثير في الحالة الأولى (تقليص البيانات) التي صممت من أجلها. كما يمكن استخدام تحليل المكوّن المستقل في النوع الأول من الحالة، ولكنه صمم للغاية الثانية (إشارات غير ممزوجة)، وهو حالياً تقنيتها الرائدة.

الجدول رقم 5.8: نتائج تحليل المكوّن الرئيسي، والانحدار، مستخدمين مكونات رئيسة، في بيانات انتخابات 2008 على مستوى المحافظة.

	تحليل المكوّن الرئيسي					انحدار مكونات حصة التصويت	
	متغيرات ذات تحميلات إيجابية	متغيرات ذات تحميلات سلبية	٪ التباين في المتغيرات المشاركة	القيم الذاتية	ارتباط ذو متغيرين بحصة أوباما من الأصوات	معامل خطأ المعيار	٪ التباين الموضحة
المكوّن 1	٪ < تعليم ثانوي، ٪ نسبة الفقر	متوسط الدخل، ٪ وظيفي/ إداري	25.31	5.56	0.089	0.560***(0.084)	0.76
المكوّن 2	٪ المهاجر، الكثافة السكانية	٪ العمر +65، ٪ فقط مستوى ثانوي	21.59	4.74	0.463	3.152***(0.091)	21.44
المكوّن 3	٪ المطلق، ٪ السود	٪ العمر < 18، ٪ لاتيني / a	9.79	2.15	0.343	3.467***(0.136)	11.75
المكوّن 4	٪ العمر 18 >، حجم الأسرة	٪ العمر 65+، ٪ لاتيني/ a	6.96	1.53	-0.071	-0.859***(0.161)	0.48
المكوّن 5	٪ الأزواج من نفس الجنس	٪ الشباب، ٪ السود	4.81	1.05	0.073	1.052***(0.194)	0.53
المكوّن 6	٪ الشباب، ٪ البيض	٪ السود، ٪ العمر 65+	4.40	0.96	-0.216	-3.267***(0.203)	4.67
المكوّن 7	٪ البروتستانت الإنجيلي، ٪ درجة التعليم الثانوي بالضبط	البطالة بين الرجال، البطالة بين النساء	4.14	0.91	-0.189	-2.940***(0.209)	3.58

***p<.001

لقد صمم تحليل المكوّن المستقل باعتباره طريقة من طرق فصل المصدر الأعمى (Blind Source Separation)، الذي يعد «مشكلة حفل الكوكتيل»، حالتها الكلاسيكية. ودعنا نقل إن لدينا ثلاثة أشخاص يتحدثون في حفل كوكتيل، ونقوم بتسجيل محادثاتهم باستعمال ثلاث ميكروفونات موضوعة عشوائياً في الغرفة. كُلّ ميكرفون سيولّد تسجيلاً يُعد مزيجاً من محادثة المتحدثين الثلاثة، ونريد طريقة تفصل التسجيلات الثلاث، كي يتسنى لنا فصل - قدر الإمكان - صوت كُلّ فرد على حدة. وفي هذه الحالة يمكننا القيام بافتراض حاسم للاستقلال الإحصائي للموجات الصوتية المنبعثة من المتحدثين الثلاثة. إن الرفع من قوة هذا الاستقلال المفترض، يسمح لتحليل المكوّن المستقل أن يُنجز بشكل رائع، لهذا النوع من مشكلة التصنيف.

ويتميز تحليل المكوّن المستقل، إذن، عن تحليل المكوّن الرئيسي من خلال استخدام استقلال إحصائه - عوض عدم ارتباطية (Uncorrelatedness) - باعتباره مبدأً موجهاً من أجل فصل البيانات إلى مكونات. ولكن كيف يختلف الاستقلال وعدم الارتباطية؟ إن الاستقلال في الأساس حالة أقوى بكثير؛ فلكي يكون متغيران غير مرتبطين، يقتضي ذلك فقط عدم توافرهما على علاقة خطية (Linear) فيما بينهما. ومع ذلك، قد تكون لديهما علاقة لا خطية مميزة. إن التعامد (Orthogonality) أو عدم الارتباط، حالة ضرورية ولكن غير كافية بالنسبة إلى الاستقلال.

والآن إذا كان متغيران غير مرتبطين، وموزعين بشكل عادي، فسيكونان - أصلاً - مستقلين. وبما أن تحليل المكوّن الرئيسي، يستخرج مكونات غوسية (Gaussian)، فإن الفرق بين عدم الارتباطية والاستقلال - بالنسبة إلى تحليل المكوّن الرئيسي - هو أمر خلافي. ومع ذلك، يفترض تحليل المكوّن المستقل، تكون البنية الرئيسة للبيانات من عناصر لا غوسية (Non-Gaussian). وقد تمّ وصف تحليل المكوّن المستقل - في واقع الأمر - باعتباره تحليل عامل لا غوسي.

هذه نقطة مهمة، وجب التركيز عليها بالنسبة إلى مستخدمي تحليل المكوّن المستقل. ويجب استخدام كُلّ الطرق في حالات تكون فيها مناسبة للمهمة القائمة، وبالنسبة إلى كُلّ جزء من الحالة إذا كانت الافتراضات التي توجه المنهجية تبدو معقولة بالنسبة إلى حالة العالم الحقيقي الذي نحن بصدد تحليله. إن لدى استخدام تحليل

المكوّن المستقل معنى، إذا وفقط إذا كنا نظن ان المكوّنات الرئيسة للبيانات مستمرة، ولكن غير موزعة - في الحقيقة - بشكل عادي إذا كانت لا غوسية بالحد الأقصى. وعملياً، ترتبط اللا غوسية (Non-Gaussianity) بالتفرطح (Kurtosis) - «بلوغ ذروة» توزيع المكوّنات قيد الدراسة. وإن تحليل المكوّن المستقل، تستخرج المكوّنات التي بلغت الذروة بشكل كبير (Leptokurtic)، أو لم تبلغها للغاية (Platykurtic). ولهذا، إذا كان للمرء داعٍ للاعتقاد في أن العناصر الرئيسة المؤسسة للبيانات قيد الدراسة، هي عناصر مبنية بواسطة عناصر أساسية، إما مركَّزة بشكل للغاية حول المتوسط (Mean)، وإما غير مركزة بخاصة، (أو على الأرجح، خليط من «البالكورتوز»، و«الليبتوكورتيك»)، فسيكون هذه الحالة تحليل المكوّن المستقل مثالياً. وفي المقابل، إذا كان شخص ما مقتنعاً بأن العناصر الأساسية موزعة بشكل عادي، فمن الواجب تجنب تحليل المكوّن المستقل لصالح تحليل المكوّن الرئيسي أو تحليل العامل.

إن تحليل المكوّن المستقل يعمل تبعاً للخطوات التالية:

1. **تحديد عدد المكوّنات المستقلة الواجب استخراجها**: بالتوسل بتحليل المكوّن المستقل، يستوجب على الباحث تحديد عناصر أو أبعاد أساسية عديدة يري أنها مؤسسة للبيانات القائمة. وإن هذا التدخل من قبل الباحث، هو أكبر أهمية من تحليل المكوّن الرئيسي أو تحليل العامل. وفي هذه الحالات الأخيرة، تولّد البرامج - عادة - عوامل أو مكونات عديدة، بقدر تعدد المتغيرات المستعملة في التحليل، ويقرر الباحث بعد العملية (بعد تحليل رسم بياني ما باستخدام معايير أخرى) العدد الذي يتم الاحتفاظ به. إن تحليل المكوّن المستقل، بالمقابل، سيستخرج فقط عدد المكوّنات التي يشترطها الباحث سلفاً. وفي حالات فصل مصدر أعمى، يتم - عادة - تَعَرُّفُ مصادر الإشارة المستقلة، ومن ثم، فإن هذه المحدودية لا تطرح مشكلة. ولكن في حالات العلوم الإنسانية، حيث يكون - عادة - عدد العناصر أو المكوّنات الرئيسة غير معروف، يكون الأمر أكثر صعوبة. وإلى حدّ علمي، فإن تقابلات الرسم البياني أو نسبة التباين الموضح، لم يتم تطويرها من أجل تحليل المكوّن المستقل.

2. **تبييض البيانات**: يستمر البرنامج في إنتاج مجموعة من المكوّنات غير المترابطة كما تم في تحليل المكوّن الرئيسي.

3. إيجاد دوران فكّ الارتباط (Decorrelation) للمكونات اللا غوسية بالحد الأقصى.

يمكن تحديد اللا عيارية (Nonnormality)من خلال إحدى الطريقتين التاليتين:

- الطريقة الأولى، فتجد المكوّنات التي يتنوع تفرطحها[4] (Kurtosis) (إيجاباً أو سلباً) من خلال تفرطح توزيع عادي.
- الطريقة الثانية، فتستخدم كمية إحصائية تدعى «الأنتروبي السلبي» (Negentropy)، وتعني الفرق في الأنتروبية[5] (Entropy) المرتبطة بما يمكن توقعه في توزيع عادي ذي تباين مماثل.

مثال تحليل المكوّن المستقل باستخدام R

لقد تم دمج تحليل المكوّن المستقل في أي برنامج من برامج البرمجيات التجارية الرئيسة، مثل «الستاتا»، والحزمة الإحصائية للعلوم الاجتماعية، ونظام التحليل الإحصائي (ومع ذلك، يمكن للمرء برمجة تحليل المكوّن الرئيسي بالنسبة إلى «الستاتا» أو نظام التحليل الإحصائي، بالتوسل بقدر كافٍ من القطع الرياضية). ومع ذلك، فقد تمت كتابة بعض البرامج التي تنجز تحليل المكوّن المستقل بالنسبة إلى R (وبالنسبة إلى MATLAB مختبر المصفوفة). وهنا نبين كيفية تنفيذ تحليل المكوّن المستقل باستخدام حزمة R التي تدعى تحليل المكوّن المستقل السريع (FastICA) (Marchini, Heaton, and Ripley 2012).

أما بخصوص تحليل المكوّن الرئيسي، فنستخدم بيانات انتخابات 2012 على مستوى المحافظة. ونسترجع رزمة تحليل المكوّن المستقل السريع من R، ونقوم بتحميل البرنامج في ذاكرة التشغيل (Working Memory):

```
Install.packages (“fastICA”)

Library (“fastICA”)
```

(4) التفرطح (Kurtosis) تعني الحدّ من ذروة منحني التردد التوزيعي (المراجع).

(5) ورد مصطلح أنتروبي (Entropy) كثيراً في الكتاب وهو يعني الانخفاض التدريجي في الاضطراب، أو انعدام النظام أو إمكانية التنبؤ. وبصورة أدق، في نظرية المعلومات يعتبر الأنتروبي مقياس لوغاريثمي لمعدل نقل المعلومة في رسالة أو لغة معينة (المراجع).

وبعد ذلك، نقوم بربط أعمدة المتغيرات التي نرغب في استخراج المكوّنات المستقلة منها، ونخزنها في مصفوفة X. وهنا ننتقي 21 متغيراً منفصلاً، ونقيس خصائص ديموغرافية والمستوى الاقتصادي والاجتماعي للمحافظات، ونشكلها في مصفوفة 21×3114.

```
x<- cbind (lnpopdens, agelt18, age1834, age65over, perwhite, perasian, perblack, perlatin, edhigher, edhs, edlhs, unempmale, unempfem, perpov_q, imdens, divorce2per, samesexper, evprot10, hhsize, occprofman, medinc)
```

إن برنامج تحليل المكوّن المستقل السريع، يُنفّذُ من خلال الرمز التالي:

```
= ical<-fastICA (x, 5, alg.typ =  «parallel», fun = «logcosh», row.norm
(TRUE, maxit = 200, tol = 0.00001, verbose = TRUE)
```

ولنتعقب بالضبط، ما نحن بصدد القيام به هنا. إننا بصدد توليد شيء يدعى «icaI» من خلال إنجاز دالة تحليل المكوّن المستقل السريع على مستوى الشيء X، مصفوفتنا المكوّنة من 21 متغيراً. أما الخيار الموالي، فيخبرنا بتحليل المكوّن المستقل السريع لتوليد خمسة مكونات مستقلة (بحيث يكون العدد المختار – في هذه الحالة – عشوائياً بما أننا لا نستند إلى معرفة قبلية أو إلى نظرية ما). وبعد ذلك، انتقينا «alg.typ = «Parallel مما يعني أن البرنامج، سيستخرج المكوّنات في آنٍ واحد. وفي المقابل، إذا ما حددنا «الانكماش»، ستُستخرج المكوّنات فرادى. وليس ثمة توجيه كبير بشأن هذا القرار، وإذا ما كان المرء منشغلاً فقط بمحاولة استخراج المكوّنات المستقلة من البيانات، فلن يكون الأمر مهماً كثيراً؛ فالمكوّنات المنتقاة في تحليلنا، بالاستخراج الموازي أو التسلسلي، غير مترابطة على حدّ سواء، ومترابطة باعتدال بعضها ببعض.

ثم، هناك سلسلة من الخيارات المترابطة بسرعة التقارب. وإن تحليل المكوّن المستقل هو خوارزمية تكرارية، تبحث عن مكونات غير مترابطة لا غوسية بالحد الأقصى. ولكن هناك طريقتين مختلفتين لتعظيم اللا غوسية، المحددة عبر الخيار الممتع (Fun Option). ويمكن انتقاء سواء دالة أساسية (Fun = «exp«) أو

خوارزمية جيب التمام القطعي (Hyperbolic Cosine) (Fun = «Logcosh»). وكلاهما يعمل جيداً، بحسب مطوري تحليل المكوّن الرئيسي، ولكن «Logcosh» في تجربتنا أسرع قليلاً. وبعد ذلك، نحتاج إلى اختيار ما إن كان يستوجب على سطور مصفوفة البيانات، التطبيع قبل التحليل. وإن انتقاء TRUE، يفضي إلى التقاء أسرع قليلاً. ويراقب الخياران المواليان بشكل أكثر مباشرة، عدد التكرارات التي تحدث قبل أن يُسمح للبرنامج أن يستقر على نتائج ما.

أولاً: نختار الحدّ الأقصى من عدد التكرارات لإنجازها.

ثانياً: نختار التسامح، الذي يعد كمية إحصائية من التناسبية. وعموماً، إن الالتقاء سريع – إلى حدّ ما – مستخدمين تحليل المكوّن المستقل السريع، ومن ثم، فإننا ننصح بتحديد «الماكسيت» (Maxit) عالياً نسبياً. ويجب أن ينظر إليه باعتباره ضمانة أكثر من أي شيء آخر. وفي المقابل، على المرء مراقبة جودة الالتقاء مع مَعْلم التسامح. وستقود القيم الأعلى إلى التقاء أسرع، ولكن ستكون أقل موثوقية. ولهذا، ننصح بتحديد «التول» (tol) في مستوى منخفض. وفي بياناتنا، على الرغم من تحديد «التول» لشيء ضئيل بشكل مطلق (مثلـ tol = 0.000000000002)، فإن الالتقاء يحدث بعد 32 تكراراً فقط، ولو أن وجود مجموعة بيانات أكبر أو متغيرات مُدخل أكثر، يستغرق وقتاً أطول. **وأخيراً** – وكما هو الحال بالنسبة إلى العديد من تحكمات R – هناك خيار الفيربوز (Verbose). وإن اختيار TRUE، سيسمح لك بمعرفة عدد التكرارات التي تحدث قبل الالتقاء وما هو التسامح المحدد في كُلّ خطوة.

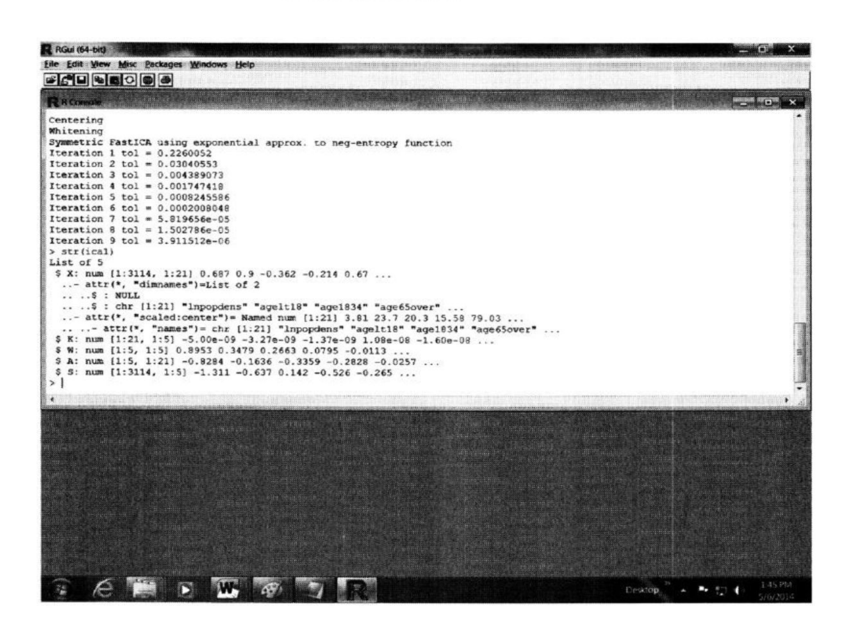
```
ical<-fastICA(X, 5, alg.typ = "parallel", fun = "exp", row.norm=FALSE, maxit=200, tol=0.00001, verbose=TRUE)
entering
hitening
ymmetric FastICA using exponential approx. to neg-entropy function
teration 1 tol = 0.2260052
teration 2 tol = 0.03040553
teration 3 tol = 0.004389073
teration 4 tol = 0.001747418
teration 5 tol = 0.0008245586
teration 6 tol = 0.0002008048
teration 7 tol = 5.819656e-05
teration 8 tol = 1.502786e-05
teration 9 tol = 3.911512e-06
```

الشكل رقم 4.8: مُخرج تحليل المكوّن المستقل في R، موضحاً التقاء النموذج.

ونشغل التحكم، ونحصل على المخرج المبين في الشكل رقم 4.8. ويطلعنا هذا فقط على أننا نحصل على الالتقاء في تسع تكرارات، ولو بتسامح منخفض. وكما هو نموذجي مع R، فإن المُخرج الآني غير مفيد. ولكن يمكن النظر إلى بناء الشيء الذي نولّده (IcaI) في الشكل رقم 5.8.

ولاحظ أن للشيء عدد من المكوّنات (تظهر هنا باعتبارها ($ X, $K, etc.) والأنسب بشكل مباشر هي أسطر تسمى A$، وأما S.SA$، فتحتوي على ترجيحات لكُلّ متغير، تُستخرج منها العوامل، ولكن النظر إليها بشكل قابل للفهم، يقتضي تحويلها.

Pairs (ica1$S)



الشكل رقم 5.8: مُخرج تحليل المكوّن المستقل في R،

موضحاً عناصر مخزنة في الشيء icaI.

إن الأسطر المبينة في المخرج R في الشكل رقم 6.8 مرتبط بمتغيراتنا الأصلية؛ أي أعمدة، مترابطة بمكونات خمسة. كما أن المكوّنات المستقلة، مثل المكوّنات الرئيسة، مشكلة باعتبارها مزيجاً خطياً (أي مجموعاً مرجحاً، تحديداً) لهذه المتغيرات. ويعد

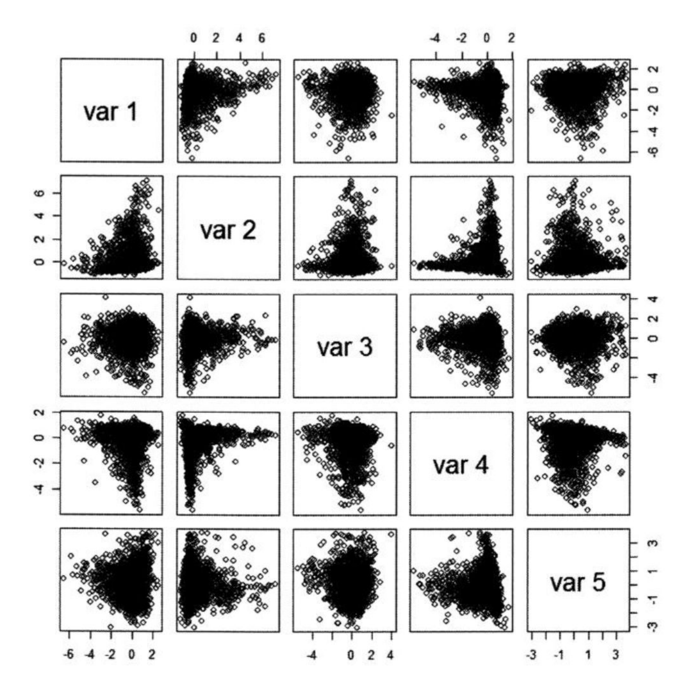
المتغير 1 للسطر هنا، الخوارزمية الطبيعية للكثافة السكانية (وإذا ما قرأنا عبر السطر الأول من الشكل رقم 6.8، فسنجد فقط تحميلات سلبية). وأما المتغير 5 للسطر، فهو نسبة سكان المحافظة من البيض غير الإسبان. وإن القراءة عبر هذا السطر يفصح عن أن هذا المتغير، مرتبط إيجاباً بالكل، ما عدا المكوّن الثاني. ويمكن إنتاج مصفوفة الرسم البياني للتشتت (الشكل رقم 7.8) التي ستبين استقلالية المكوّنات:



الشكل رقم 6.8: تحميلاً المتغيرات (سطور) على مستوى المكوّنات (الأعمدة) من تحليل المكوّن المستقل (باستخدام تحليل المكوّن المستقل السريع لحزمة R).

والآن، يمكننا استخدام هذه المتغيرات باعتبارها متنبئات في انحدار ما، متنبئين حصة أوباما من الأصوات في محافظات معينة (الجدول رقم 6.8). ويبدأ الانحدار في العمود المسمى (1) بالمكوّن المستقل الأول، وبعدها يضيف الباقي، الواحد تلو الآخر. ومن المهم التذكير بأن المتغير التابع لم يكن عضواً من مجموع المتغيرات التي استخرجت منها المكوّنات المستقلة، وبالتالي، فإن أي تباين في المتغير التابع الذي تمّ شرحه من خلال المكوّنات المستقلة، إما بسبب الصدفة العشوائية، أو بسبب

علاقته بالمتغيرات الأصلية في المجموعة. ومن المفيد أيضاً التذكير بأن نموذج انحدار ما، الذي يستخدم هذه المجموعة الكاملة من المتنبئات (أي قبل القيام بتحليل المكوّن المستقل)، كان لديه R^2 بنسبة 5855.



الشكل رقم 7.8: مصفوفة الرسم البياني لتشتت المكونات المستقلة (في R).

وتخزن القيم الحقيقية لكُلّ مكون في كُلّ حالة على حدة، في المكوّن S. وستكون أيسر – نوعاً ما – النظر في هذا، إذا حولنا S إلى سلسلة من متغيرات خمسة:

```
Comp1<-ica1$S{1 : 3114, 1}
comp2<-ica1$S {1 : 3114, 2}
comp3<-ica1$S {1 : 3114, 3}
comp4<-ica1$S {1 : 3114, 4}
comp5<-ica1$S {1 : 3114, 5}
```

الجدول رقم 6.8: انحدار حصة أوباما

من الأصوات على مستوى المكوّنات من تحليل المكوّن المستقل.

	(1)	(2)	(3)	(4)	(5)
المكوّن 1	-5.0855***	-5.085***	-5.085 ***	-5.085 ***	-5.085***
	(0.2496)	(0.238)	(0.238)	(0.215)	(0.214)
المكوّن 2	-	4.214***	4.213***	4.213 ***	4.213 ***
		(0.238)	(0.238)	(0.215)	(0.214)
المكوّن 3	-	-	(0.238)	(0.215)	(0.214)
			0.035	0.035	0.035
المكوّن 4	-	-	-	5.667***	5.667***
				(0.215)	(0.214)
المكوّن 5	-	-	-		-1.255***
					(0.214)
ثابت	(0.2496)	(0.238)	(0.238)	(0.215)	(0.214)
	38.4431	38.443	38.443	38.443	38.443
R^2	1177.	1984.	1984.	3446.	3517.
Adj.R^2	1174.	1979.	1977.	3437.	3507.

*** p<.001.

وتظهر بعض الأشياء مباشرة من التحليل

أولاً: إن المجموعة الكاملة للمكونات المستقلة الخمسة تمتلك ما هو أكثر من نصف القوة التوضيحية لنموذج المتغير الأصلي. وهكذا، على الرغم من أننا أنجزنا تبسيطاً كبيراً للبيانات، فإن ذلك تم على حساب تخفيض معتبر للقوة التنبؤية لنموذجنا.

ثانياً: إن مسألة أن المتغيرات غير مترابطة فيما بينها، تم إظهارها مباشرة من خلال كون - كما هو الحال بالنسبة إلى تحليل المكوّن الرئيسي - معاملات الانحدار، لا تتغير عند إضافة مكونات إضافية. وفي الحقيقة، إذا تغيرت فعلاً، فسيتضمن ذلك ارتباطاً كبيراً بين المتغيرات. ولكن الانحدار الخطي - بطبيعة الحال - يمكنه فقط تقديم تلميحات حول ما إذا كانت المكوّنات مترابطة أم غير ذلك، وليس ما إذا كانت مستقلة.

ثالثاً: ثمة شيء غريب يحدث للأخطاء المعيارية في هذا التحليل؛ فهي نفسها بالضبط بالنسبة إلى كُلّ المكوّنات (والاعتراض) داخل النماذج. وهي أيضاً متشابهة عبر النماذج كلها. وهذا ربما راجع إلى أن المكوّنات لا غوسية بالحد الأقصى. وباعتبارها متغيرات غير طبيعية (التي تعد غير طبيعية بشكل متشابه)، فهي تخلط بالضبط محاولة حساب خطأ معياري، الذي يفترض المعيارية. **وأخيراً**، إن المتغير الثابت، لا يتغير تماماً انطلاقاً من نموذج إلى آخر. وهي – في الواقع – متساوية مع متوسط المتغير التابع بالنسبة إلى البيانات جميعها. (نعم إن متوسط الأصوات بحسب المحافظات كانت 38.44٪. وقد فاز أوباما بمهارة في المحافظات الأكثر كثافة سكانية، وخسر المحافظات ذات كثافة سكانية قليلة، والسكان متمركزون في عدد صغير نسبياً من المحافظات). وهذا راجع إلى كون البيانات قد تم تعييرها قبل استخراجنا للمكونات. وبالتالي، إن كُلّ المكوّنات لها متوسطات قريبة جداً من الصفر. وكون أن المتغير الثابت يبقى في متوسط السكان، فإن ذلك يعني أنه صحيح أصلاً.

خلاصة

تستخدم طرق استخراج المتغير لتقليص عدد المتغيرات قبل مباشرة التحليل، عبر استكشاف عدد صغير من مكونات عوامل غير مترابطة، تلخص عدداً أكبر من متغيرات مقاسة. وفي مقابل طرق تقليص البيانات التي تمت مناقشتها سابقاً (مثل الانحدار التدريجي)، التي تنتقي المتنبئات الأكثر أهمية من بين قائمة أطول لمتنبئات المرشَّح، تحاول طرق استخراج المتغير تلخيص جميع المتغيرات المتاحة. وبشكل أدق، يقوم استخراج المتغير بعملية تحليل مصفوفة التغاير التي تصف العلاقات بين المتغيرات المقاسة. إن تحليل المكوّن الرئيسي، وتحليل المكوّن المستقل كلاهما، يحاولان تلخيص مصفوفة التغاير ذات متجهات ذاتية عمودية قليلة أو مكوّنات قليلة.

وهذه طرق غير خاضعة للرقابة والإشراف: إذ لا يشركون متغيراً تابعاً، بل يلخصون – عوضاً عن ذلك – العلاقات بين الخصائص، أو المتنبئات، أو المتغيرات المستقلة. ولسوء الحظ، هناك مقايضة بين التبسيط والدقة. وقلّما تفسر المكوّنات المستخرجة عموم التغاير المجسد في العدد الأكبر للمتغيرات المقاسة. علاوة على

ذلك، حتى عندما تلخص المكوّنات المستخرجة معظم التغاير بين المتنبئات، لا يترتب عن ذلك كون أن النموذج يستخدم تلك المكوّنات المستخرجة للتنبؤ بمتغير تابع، سيمنح بالضرورة تنبؤات جيدة. وفي الحقيقة، وجدنا مراراً وتكراراً - في الأمثلة المبينة أعلاه - أن المتغيرات الأصلية التي تم اتّخاذها كمجموعة، كان أداؤها أفضل من حيث تنبؤ متغير تابع، من المكوّنات المستخرجة منها. ومع ذلك، تستعمل تقنيات كُلّ من تحليل المكوّن الرئيسي، وتحليل المكوّن المستقل، من قبل المختصين في التنقيب في البيانات، خاصة في حالات حيث وجود عدد كبير جداً من متغيرات مقاسة (P كبيرة جداً) حتى أن المحلل يجد نفسه مضطراً إلى تلخيصها أمام قلة الخيارات، من خلال استخراج عدد أصغر من المكوّنات. وقد يصف مختصو التنقيب في البيانات عملية الاستخراج هذه بتقليص أبعاد البيانات، مع الحفاظ على بنيتها أو نمطها الأصلي في الوقت نفسه.

الفصل التاسع
المصنفات

تعد مصنفات التنقيب في البيانات، برامج تتنبأ بفئة أو بصنف متغير تابع ما، تُصنف ضمنه ترصدات فردية. على سبيل المثال، قمنا سابقاً بتصنيف الأفراد وفق توافرهم على تأمين صحي من عدمه، متوسلين بعدد من الخصائص الديموغرافية. وفي بعض أنواع خوارزميات التصنيف حيث يشمل التصنيف تطوير نموذج إحصائي تنبؤي، من خلال استخدام مجموعة من متغيرات مستقلة، أو خصائص تتنبأ بقيمة كُلّ فرد على مستوى نتيجة متغير تابع أو هدف. ويستخدم ذلك التنبؤ - الذي يتمثل في شكل احتمالية تصنيف حالة معينة ما، ضمن فئة أو صنف معين - لتصنيف الفئة التي سيُخصص لها ترصد معين.

أما بعض الأنواع الأخرى من خوارزمية التصنيف، فلا تستخدم نموذج تنبؤي من هذا النوع، وإنما تستخدم الطرق اللا مَعْلمية (Non-Parametric)، للبث في صنف متغير نتيجة ما، يُصنَّف ضمنه كُلّ ترصد. ولكن يشمل كُلّ تصنيف تعليماً تحت الإشراف (Supervised Learning): من خلال استخدام مجموعة بيانات تدريب، تضم حالات، يعرف الشخص من أجلها التصنيف الصحيح لكُلّ ترصد على حدة، بغية تطوير نوع من أنواع قاعدة تنبؤية. ويمكن تطبيق تلك القاعدة على مجموعة بيانات حيث لا يعرف الشخص فئة أو صنف كُلّ حالة، كي نصنف هذه الحالات الجديدة.

وفي الأقسام الآتية، نقدم أمثلة من مصنفات مختلفة عديدة، مستخدمة من قبل مختصين في التنقيب في البيانات. فقد قام علماء الحاسوب بتطوير العديد من الخوارزميات من أجل عملية التصنيف، بحيث تختلف هذه الخوارزميات تبعاً لسرعة عمليتها ودقتها. علاوة على ذلك، تعمل بعض الخوارزميات أفضل بالنسبة إلى مجموعات بيانات معينة مقارنة بأخرى. كما أن الممارسين لا يعرفون عادة – وبشكل مسبق – أي نوع من المصنف، الذي سيعمل على النحو الأفضل بالنسبة إلى بياناتهم، ومن ثم، لا غرو أن يتم تجريب عدة مصنفات ومقارنة دقتها على مستوى اختبار مجموعات البيانات، أو حتى الجمع بين التنبؤات المستمدة من هذه المصنفات المختلفة في مجموعة واحدة، فيما أصبح يعرف بعملية التعلم بالمجموعة. وغالباً ما تسفر عملية الجمع بين مصنفات مختلفة في مجموعة واحدة عن نتيجة أكثر دقة مقارنة مع أفضل المصنفات الفردية.

k- أقرب الجيران

إن مصنف k– أقرب الجيران (KNN)، طريقة تصنيف لا مَعْلمية، وباعتباره مصنفاً من المصنفات، فهو بسيط جداً وبديهي. تصور أن لدينا مجموعة S من نقطة بيانات، نود تقدير عضويتها ضمن فئة من أصل اثنتين. ولدينا معلومات عن قيمة هذه النقاط على مستوى متغيرات أخرى، X. وهذا يعني – من بين أشياء أخرى – إمكانية تحديد موقع كُلّ من نقاط البيانات في S في حيز متعدد الأبعاد، المحدد من قبل هذه المتغيرات المدخلة لـ X . ويمكن تحديد أقرب – أقرب الجيران لكُلّ عضو من S من حيث وجود قيم مماثلة على مستوى X – من بين نقاط البيانات الأخرى. وبعد ذلك يمكننا تخصيص كُلّ نقطة بيانات Si إلى الفئة التي ينتمي إليها معظم أقرب جيرانها.

فعلى سبيل المثال، قد تكون لدينا بيانات تصف مجموعة مكونة من أطفال يبلغون من العمر ثلاث أو أربع سنين. وانطلاقاً من هذه البيانات، نعرف بعض الأشياء عن كُلّ الأطفال دخل أسرهم، وتحصيلهم العلمي، ومنزلة القوة العمالية، وكثافة السكان، ومتوسط الدخل المنزلي لمسالك تعدادها، وغيرها. وفي هذه الحالة، يمكننا استخدام مصنف KNN للتنبؤ بوضعية طفل ما قبل المدرسة، وذلك ببساطة عن طريق تخصيص لذلك الطفل، وضعية ما قبل المدرسة للأطفال الآخرين الذين

يعدون أكثر مماثلة له من حيث قياسات تدابير الأسرة والجوار. وفي الأساس، إن ما نقوم به في هذه التقنية، هو أخذ حالة، والبحث من حولها في حالات أخر، مشابهة، واستخدام هذه الحالات لتخمين عضوية الفئة المنتمية إليها.

ويمكن استخدام هذه التقنية للقيام بأكثر من مجرد تصنيف ثنائي؛ إذ بالإمكان استخدامها أيضاً من أجل تصنيف متعدد الفئات. (على الرغم من أن احتمالية حدوث «تعادل» يزداد مع عدد الفئات)، أو من أجل تنبؤ قيمة نتيجة مستمرة. وفي هذه الحالة الأخيرة، تقوم بحساب قياس المركزية انطلاقاً من الجيران) الذين يعدون - وعلى نحو أكثر شيوعاً - الوسيلة أو الوسيطة) وتطبيقها باعتبارها تنبؤاً للحالة قيد الدراسة. ومن ثم، فإن انحدار KNN مماثل تماماً لتقنيات التمهيد المحلي، القائم على النواة مثل انحدار خطي محلي (Altman, 1992).

هناك بعض التساؤلات الأولية التي ستصادف المرء قبل أداء هذه التقنية:

أولاً: كم عدد الجيران الذين يستوجب على المرء اختيارهم؟ يمكن لهذا الاختيار أن تنتج عنه نتائج هامة، على خلفية إمكانية تخصيص حالات إلى فئات متعددة استناداً إلى ما إن تم - مثلاً - «إحصاء» ثلاثة من أقرب الجيران، عوض سبعة منها. وفي صياغة سابقة، رأى كُلّ من كوفر وهارت (Cover and Hart 1967) أن استخدام جار واحد يمكن أن يكون كافياً، أو أفضل أحياناً. ومع ذلك، اقترح هاستي (Hastie) وتيبشيراني (Tibshirani) (1996) استناد مثالية جار واحد - بشكل كبير - إلى عدد السمات المستخدمة لتحديد المسافة. وإن المساحة الدائرية المستخدمة للبحث عن حالة ما، تزداد مع ازدياد عدد المتنبئات المستخدمة، وذلك بجذب مزيد من الحالات البعيدة إلى أقرب الجيران قدر الإمكان.

ويشمل حلّ إشكالية عدد الجيران المستخدمة - ودون غرابة - تقنية لا مَعْلمية أخرى. ويمكن للمرئ استخدام الصلاحية المتبادلة لانتقاء أفضل قيمة لـ k. وتحديداً، يمكننا تقسيم البيانات عشوائياً إلى ثلاثة أجزاء:

- التدريب.
- الصلاحية.
- بيانات الاختبار.

ونقوم بتوليد تقديرات باستخدام عدة قيم مختلفة من k في مجموعة التدريب، ثم نقوم بانتقاء أفضل قيمة لـ k باستعمال مجموعة الصلاحية لمعرفة نوع k الذي ينتج تصنيفاً أكثر دقة. **وأخيراً**، نقوم بتقييم التناسب في مجموعة بيانات الاختبار.

أما المسألة التمهيدية الثانية، فتتجلى في تحديد المعيار، للبتّ في نقاط البيانات الأكثر قرباً؛ أي ما هو نوع المسافة التي سيستخدمها المرء في تحديد النقاط «الأكثر قرباً». ومن المألوف جداً أن تستخدم تقنيات KNN مسافة أقليدية، أو مسافة «مانهاتن» (مجمع المدينة) أو مسافة مالينوفسكي، ولو أنه يمكن استخدام أنواع أخرى من المسافة (ماهالانوبيس، على سبيل المثال).

أما المسألة الثالثة - وفي علاقة بالمسألتين السابقتين - فهي تهم «عملية فرز الأصوات»؛ أي إنه، بعد اختيار k، وتحديد كيفية قياس المسافة، سنحصل بالنسبة إلى كُلّ نقطة من نقاط البيانات الأخرى، على مجموعة بيانات أخرى لـ *k*، التي تقدم معلومات من أجل تنبؤ التصنيف. وإن نقاط k هي في الأساس، «التصويت» على العضوية أو الصنف للحالة المستهدفة. ولكن، بما أن هذه النقاط من نقاط *k* قد لا تتفق، فكيف يجب علينا عدّ هذه الأصوات؟ فهل ينبغي عدها جميعاً على قدم المساواة؟ أم يجب علينا اعتبار نقاط البيانات الأقرب أكثر إفادة؟ عموماً، ينبغي ممارسة التمرين من خلال ترجيح الأصوات عكسياً للمسافة انطلاقاً من الترصد المعني بالدراسة (Dudani, 1976). وبالقيام بهذا، بشكل عرضي، يخفف إلى حدّ ما من تبعات اختيار *k*، أي إنه لما نزيد من قيمة *k*، فإننا نقوم بالزيادة في حجم الحيز حول نقطة البيانات التي نبحث من خلالها عن معلومات حول عضوية الصنف، وبالقيام بذلك نزيد من احتمال ارتكابنا لخطأ ما، لأنه يمكننا «العبور» من حيز ما، حيث هيمنة فئة واحدة، إلى حيز حيث هيمنة الصنف الآخر. ويعد هذا مهماً خاصةً، بالنسبة إلى حالات الحدود) أي إن الحالات في فئة واحدة الأكثر تماثلاً لحالات في الفئة الأخرى). ولكن الترجيح بواسطة مسافة عكسية، يقلل من أهمية الحالات الأكثر بعداً، ويزيد من تأثير الحالات الأكثر قرباً.

وثمة مسألة تمهيدية أخيرة مهمة، تتجلى في عدد متغيرات المتنبئ التي تستخدم في تحديد المسافة - الـ X التي نوقشت سابقاً. وعلى ما يبدو - وعلى نحو حدسي - إن اختيار أكبر عدد ممكن من المتغيرات، قد يكون مثالياً، بما أن ذلك من شأنه أن يزودنا بمزيد من المعلومات التي تهم الحالات التي تعد «فعلاً» مماثلة، عوض فقط كونها متماثلة على مستوى عدد قليلٍ من خصائص مختارة عشوائياً للغاية. ومع ذلك، إن الذي عرض، هو إمكانية أن يطرح وجود قدر كبير من المعلومات، مشكلة. وإن زيادة عدد السمات أو المتنبئات، يزيد من أبعاد حيز البحث، ومن ثم، الحجم العام لحيز البحث (فكر في الانتقال من دائرة تحيط بنقطة ما إلى مجال ذي شعاع Radius تلك النقطة). وبقيامنا بذلك، ينتهي بنا المطاف إلى زيادة عدد «الجيران» المتساوية الأبعاد (Equidistant) انطلاقاً من النقطة قيد الدراسة (أي تلك التي نريد تصنيفها). ومن خلال عدد كافٍ من السمات، ننتهي بحيز بحث، تم وصفه من قبل حيز-n (أي حيز في أبعاد n، حيث إن n يشكل عدد السمات) الذي يشكل سطحه عدداً كبيراً من نقاط البيانات التي «المتعادلة» من حيث المسافة من نقطة المركز. وفي هذه الحالة، يتم تسوية طريقة k-أقرب الجيران على نحو حتمي من قبل لعنة البعدية (Tibshirani and Hastie) (1996).

ويقتضي وجود عدد كبير من السمات - إذن - طريقة من طرق تخفيض البعدية - سواء كان ذلك باستخراج السمة أو انتقائها (أو حتى الجمع بين الاثنين). كما يمكن استخدام المكوّنات الرئيسة أو الإسقاط العشوائي لطي أبعاد الحيز؛ أو يمكننا رسم «اللاسو» (Lasso) تدريجياً، أو انتقاء الأبعاد الأكثر أهمية باعتماد المراحل.

ولم يجد k-أقرب الجيران أبداً مأوى له في العلوم الاجتماعية على الرغم من حضوره في أشكال مختلفة منذ عقود، (باستثناء حالة واحدة، انظر (Qian 2010))، واستعماله في إعدادات تطبيقية مثل إدراك الوجه، وتصنيف النصوص، والبيولوجيا، وفحص تطبيق الائتمان.

k-أقرب الجيران باستخدام منمذِج الحزمة الإحصائية للعلوم الاجتماعية

لقد تمت كتابة برامج لتشغيل مصنف من مصنفات k-أقرب الجيران لدى كُلّ من المتالاب (MATLAB) وR (حزمة k-أقرب الجيران). وإن حزمة التنقيب في البيانات

للحزمة الإحصائية للعلوم الاجتماعية - المنمذِج - لديه أيضاً روتين k-أقرب الجيران، وهذا ما سنبينه أدناه، بحيث نوضح قدرته التنبؤية باستخدام بيانات من مسح المجتمع الأميركي، واستخدام k-أقرب الجيران للتنبؤ بوضع التأمين الصحي.

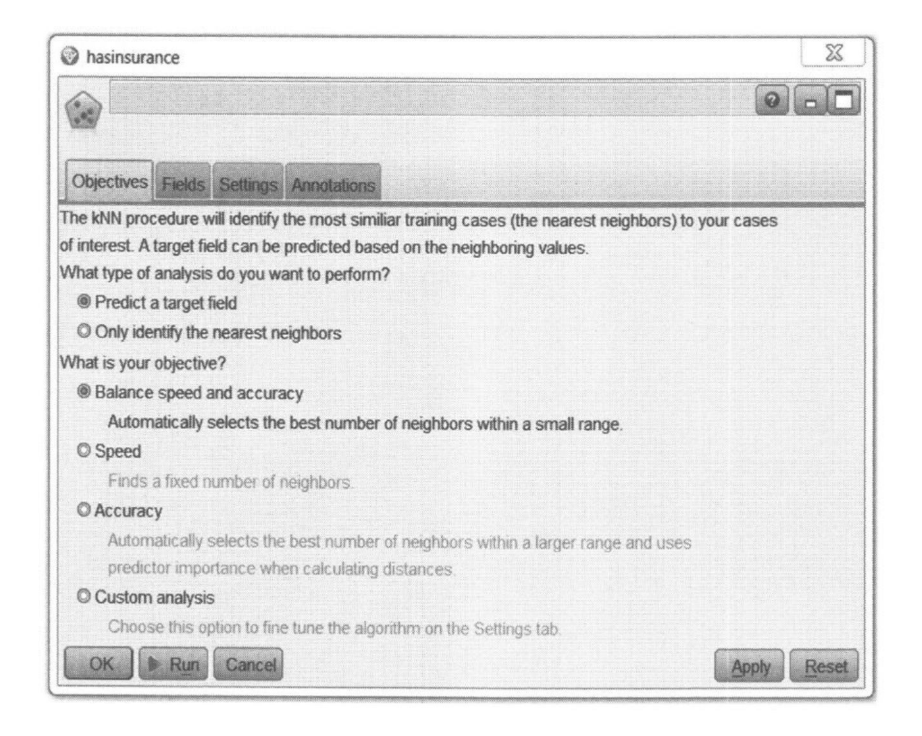
إن المنمذج نظام سهل المنال، طور لاستخدامه من قبل مختصين في التنقيب في البيانات - أشخاص في مجال الأعمال، والتسويق، وغيرهما. ومثله مثل بعض التطبيقات الحديثة الأخرى) («الرايبدماينر»، أو المنقّب السريع (RapidMiner)، مثلاً)، فإن لدى شكله «تيارات» بناء مستخدم التحليلات. ويتكون كُلّ تيار من «سلسلة عقد» متصلة، يمثل كُلّ منها سلسلة من العمليات التي طبقت على البيانات. وإن نقر أيقونة عقدة ما نقراً مزدوجاً، يفتح نافذة ذات خيارات عقدة محددة. وهذه النوافذ مماثلة تماماً لتلك الموجودة في إحصاءات الحزمة الإحصائية للعلوم الاجتماعية، كما يتم تبديل الخيارات بشكل كبير من خلال التأشير والنقر عوض الصيغة.

لقد قمنا بجمع عينة عشوائية مؤلفة من 6,000 حالة من مجموعة بياناتنا الضخمة، لأن منمذج العديد من برامج التنقيب في البيانات الأخرى، يمكن تشغيله ببطئ عند أداء عمليات معقدة، انطلاقاً من بيانات ضخمة. بالإضافة إلى هذا، قمنا بموازنة البيانات على النتيجة عند معاينتنا، لنستخلص عينة مؤلفة من 6,000 حالة مقسمة بالتساوي إلى حالات تتوافر على تأمين صحي أو لا تتوافر عليه. وقمنا بهذا لكي نزيل من البرنامج، إغراء تخصيص - ببساطة - جميع الحالات لفئة الأغلبية (وهذه استراتيجية ستسفر - على نحو عرضي - عن معدل خطأ، غير محترم يقدر بـ 13 ٪).

وفي برنامج المنمذج نقوم بتمرير ملف البيانات عبر عقدة النوع، حيث نختار المتغير الهدف ونقوم بتنظيف مستويات قياس متغيرات أخرى. وفي الخطوة التالية، نقسم البيانات إلى 50٪ من مجموعة تدريب و50٪ من مجموعة اختبار، لأن الصلاحية المتبادلة ضرورية حتماً لاستخدام مصنف k-أقرب الجيران. **وأخيراً،** نقوم بوضع عقدة k-أقرب الجيران في هذا التيار (الشكل رقم 1.9).

ولدى روتين k-أقرب الجيران عدد كبير تقريباً من الخيارات المبنية داخله، مما

يمنحه - إلى حدّ ما - قدراً كبيراً من المرونة. وبعد نقر العقدة نقراً مزدوجاً، تُفتَح نافدة، تسمح لك باختيار، في جدولة الأهداف (Objectives Tab)، ما إذا كنت تريد استخدام k-أقرب الجيران فقط للعثور على أقرب الجيران لكُلّ حالة على حدة، أم كنت تريد استخدامه باعتباره مصنفاً حقيقياً. وبما أننا نريد الاستخدام الأخير، ننتقي «تنبأ مجال هدف ما». وبعد ذلك، يطلب البرنامج ما إذا كنا حرصين على إنجاز المشروع سريعاً وبدقة متناهية، أو الجمع بين الاثنين، أو ما إن كنا نريد تخصيص النموذج. إن الخيارات الثلاثة الأولى تسمح للمستخدم بثلاث طرق مختلفة بالنسبة إلى نموذج اختيار الإعدادات الافتراضية. ونحن نشجع المستخدمين بقوة لنقر - ببساطة - «تحليل مخصص» (Custom Analysis)، والانتقال إلى الإعدادات نفسها.



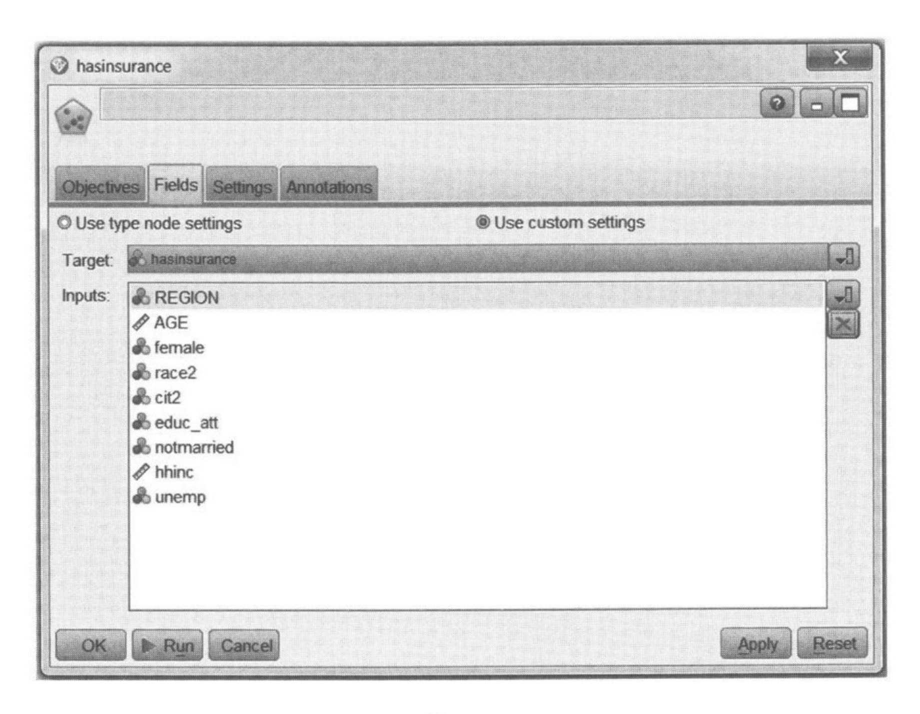
الشكل رقم 1.9: مصنف k-أقرب الجيران

في منمذج الحزمة الإحصائية للعلوم الاجتماعية.

بعد ذلك، قمنا باختيار نموذجاً في جدولة المجالات (Fields Tab) (الشكل رقم 2.9). وهنا نتوقع تغطية التأمين الصحي باستخدام منطقة التعداد، والعمر، والعرق، والنوع، والمواطنة، والتحصيل العلمي، والحالة الاجتماعية، ودخل الأسرة، والحالة الوظيفية، باعتبارها متنبئات.

وفي جدولة الإعدادات (الشكل رقم 3.9)، هناك عدد من الأقسام الفرعية التي تتيح إعداداً مَعْلمياً. وفي إطار النموذج، تختار ما إذا كنت تريد استخدام البيانات المقسمة للتحقق من صحة النتائج، ووضع متغير تقسيم خاص بك، كما يمكنك أيضاً اختيار ما إذا كنت ترغب في بناء نماذج منفصلة بالنسبة إلى مجموعات مختلفة من الحالات. وهذا يعني أن بإمكانك إدارة التصنيفات منفصلة للرجال والنساء، على سبيل المثال، أو تقسيم البيانات إلى مجموعات فرعية عشوائية وتشغيل روتينات التصنيفات المنفصلة لكُلّ واحد منها. وهذا الخيار الأخير، سيزيد بشكل كبير من مقدار الوقت الذي تستغرقه من أجل تشغيل العملية. وقد تريد ببساطة، تشغيل روتينات منفصلة بطريقة يدوية على مجموعة بيانات منفصلة، ما دام بإمكان منمذج ما حيازة أي عدد من مجموعات بيانات «مفتوحة» في آن واحد.

بعد ذلك، وتحت خانة الجيران، نقوم بوضع قيم لـ k. وإن الطريق الأسرع، هو تزويد البرنامج بـ k ثابت، ولكن بالإمكان اختيار مجال ما، وسيقوم البرنامج باختيار قيمة، تقلص من معدل التحقق من الخطأ. ويتضمن هذا عملية تشغيل تحاليل متعددة k–أقرب الجيران، مما يزيد من وقت التشغيل بشكل كبير. ومع ذاك، من الأهمية الحصول على k صحيحة، كما أن انتقاء قيمة، إما عالية أو منخفضة للغاية، سينقص من الدقة التنبؤية للبرنامج. ونقوم بوضع الحدّ الأدنى إلى 3، والحدّ الأعلى إلى 25 لتمكين البرنامج من مقدار من المرونة في اختيار القيمة الأفضل لـ k.

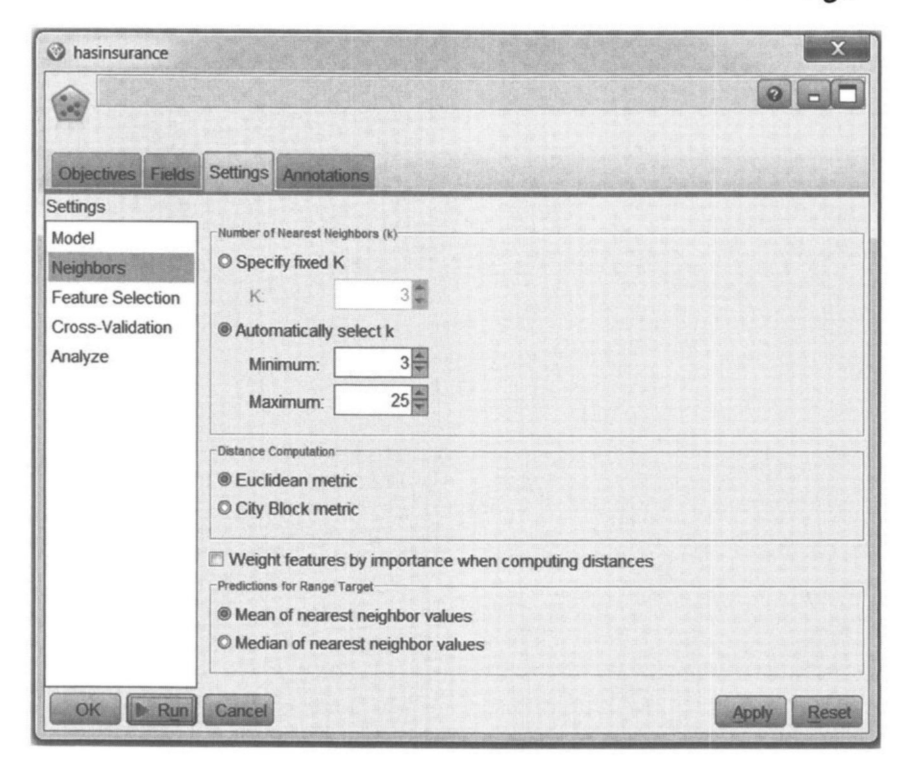


الشكل رقم 2.9: مدخل مصنف k-أقرب الجيران في منمذج الحزمة الإحصائية للعلوم الاجتماعية.

وفي هذه الجدولة، نقوم أيضاً باختيار المسافة القياسية التي سنستخدمها، وما إن تم ترجيح السمات أم لم يتم. كما يمكن للبرنامج حساب المسافة الإقليدية أو مسافة مجمع المدينة («مانهاتن»)، ونحن نفضل المسافة الإقليدية. كما نختار ترجيح المتنبئات من خلال أهميتها، متوخين بذلك اعتمادنا أكثر، على المتنبئات الأكثر أهمية في تنبؤ عضوية المجموعة - عموماً - في حساب المسافات لجيران محتملين.

ثم تختار - تحت انتقاء سمة - ما إن كنت تريد البرنامج لانتقاء السمات للبت في متغيرات المتنبئ المعينة المستخدمة. وإذا كان لديك عدد متوسط من السمات - 15 أو 20، احتمالاً - فإن استخدام طريقة ما لإزالة السمات الزائدة أو غير المفيدة، هي فكرة جيدة، على ما يبدو. ولا نواجه هذه الحالة، ونفضل عدم أداء انتقاء سمة.

تضم إعدادات الصلاحية المتبادلة مَعْلمات، يمكن تغييرها فقط في حالة عدم أداء انتقاء سمة. فهي تسمح للباحث بأداء الصلاحية المتبادلة لطية k-، وضبط نواة لتخصيص حالات بطريقة عشوائية للطيات، وذلك حتى يكون من الممكن تكرار التحليل. ونحن بصدد استخدام الصلاحية المتبادلة الكابحة، لأن ذلك أمراً غير ضروري (وفي الواقع، إن البرنامج لن يسمح لك باختيار الصلاحية المتبادلة لطية k- إذا سبق لك أن أدخلت متغيراً، لصلاحية ما). **وأخيراً**، نقوم بتشغيل النموذج بنقر التشغيل (Run).



الشكل رقم 3.9: تحديد المعلم بالنسبة إلى مصنف k-أقرب الجيران في منمذج الحزمة الإحصائية للعلوم الاجتماعية.

يظهر لنا كتلة النموذج (Model Nugget) التي أنتجها المنمذِج كيف أن معدل الخطأ تنوع مع k (الشكل رقم 4.9)، وبدأت مرتفعة نسبياً، أي حوالي 30.0٪ ، عندما كانت k 1 أو 2، وانخفضت - بسرعة في البداية، وبعد ذلك انخفضت على نحو أكثر

بطئاً - حتى بلغت k 20 (حوالي 27.5٪). وبعد هذه المرحلة، بدأ معدل الخطأ في الارتفاع مجدداً. وثمة نقطتان هامتان نشير إليهما في هذا المثال.

أولاً: يظهر معدل الخطأ علاقة خط منحني مع *k*. ويعد ارتفاع معدل الخطأ بشكل مطرد نسبياً بعد *k*=13، مهماً، لأنه يمكننا بمعرفة عدم قيامنا ربما بمجرد تعريف حدّ أدنى محلي.

ثانياً: ينبغي الإشارة إلى أن معدل الخطأ يتنوع إلى حدّ ما، ولكن ليس على نحو كبير. وربما يكون الأمر على هذا النحو بالنسبة إلى معظم البيانات، مما يشير إلى أن الأخطاء في اختيار *k* ليست بالضرورة ذات عواقب وخيمة عملياً. وهنا يظهر أن نطاق معدل الخطأ أقل من 4 نقطة في المائة. ومن ناحية أخرى، هذا يظهر فعلاً، أنه من خلال اختيار مجموعة واسعة من القيم الممكنة لـ *k*، غالباً ما يكون بالإمكان القيام بعمل أفضل من حيث التنبؤ.

وبإضافة عقدة تحليل ما إلى التيار، يمكننا فحص مدى فاعلية النموذج. ويؤدي مصنف k-أقرب الجيران بشكل باهر، من خلال تصنيف - وبشكل صحيح -74 ٪ من بيانات التدريب و75٪ من بيانات الاختبار. كما نلاحظ أيضاً قدرته التنبؤية اللائقة بالنسبة إلى كُلّ من الإيجابيات الصادقة (الذين يتوافرون على تأمين)، والسلبيات الصادقة. وفي بيانات التدريب، تبلغ نسبة نموذج الحساسية 71.5٪، والخصوصية 76.2٪. وأما الأعداد المقارنة بالنسبة إلى بيانات الاختبار، فهي 70.6٪، و 78.1٪.

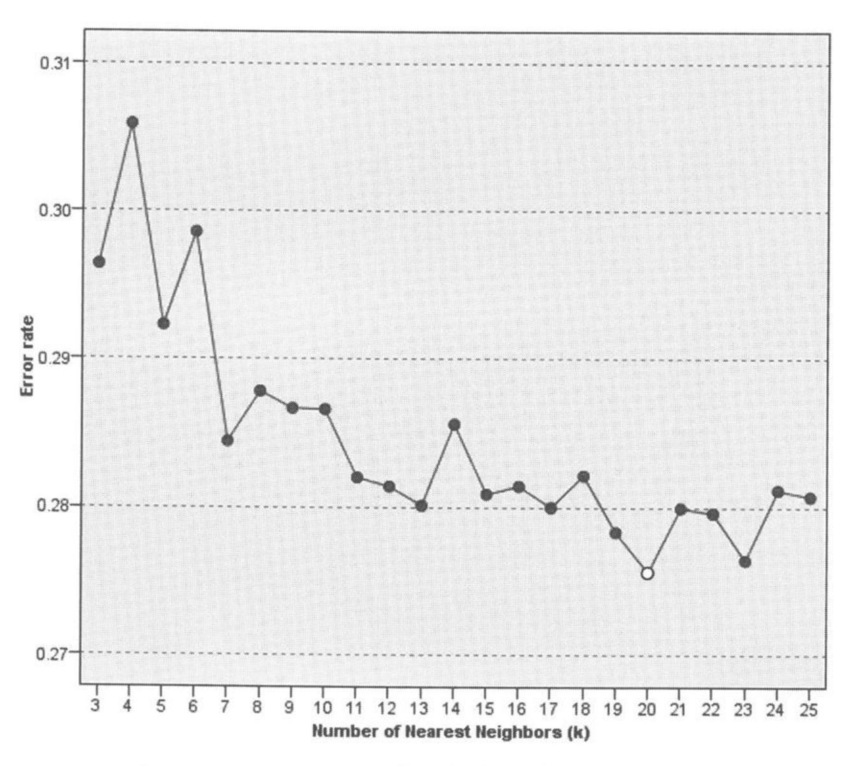
ويقوم المنمذج بحساب «درجات الميل» بالنسبة إلى تصنيفاته التي تشير إلى مدى يقين البرنامج من تنبؤه. كما نرى في تقرير عن قيم الثقة (الشكل رقم 5.9) أن المنمذج صحيح بنسبة 100٪ من الحالات، في حين إن لديه نسبة يقين من تنبؤه، تصل إلى 90.9٪ في كُلّ من بيانات التدريب وبيانات الاختبار.

كيف السبيل إلى مقارنة k-أقرب الجيران بنماذج تنبؤية أخرى في بياناتنا؟ يقارن الجدول رقم 1.9 أربع طرق أخرى:

- الانحدار اللوجيستي.
- أشجار التقسيم.

• شعاع الدعم الآلي.
• شبكة محايدة.

وتبدو الإجابة: جيدة إلى حدّ ما. وتقوم النماذج الأخرى بعمل جيد في إيجابيات صادقة تنبؤية، في حين تعمل k-أقرب الجيران عملاً أفضل في السلبيات الصادقة التنبؤية.



الشكل رقم 4.9: رسم بياني لمعدل الخطأ بقيمة k في تصنيف k-أقرب الجيران (منمذِج الحزمة الإحصائية للعلوم الاجتماعية).

لقد بدأنا مناقشتنا لأدوات التصنيف - وهو حقل ضخم في ميدان التنقيب في البيانات - بحديث عن مصنف k-أقرب الجيران اللا مَعْلمي، الذي يقوم بأداء جيد جداً في بيانات مسح المجتمع الأميركي في التنبؤ بتغطية التأمين الصحي، على الرغم

من أنه ليس متفوقاً على المصنفات الأخرى بشكل واضح. وسنواصل في القسم الموالي، تحليل خوارزميات تصنيف أخرى.

Results for output field hasinsurance

Comparing $KNN-hasinsurance with hasinsurance

'Partition'	1_Training		2_Testing	
Correct	2,175	73.75%	2,273	74.5%
Wrong	774	26.25%	778	25.5%
Total	2,949		3,051	

Coincidence Matrix for $KNN-hasinsurance (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	$null$
0.000000	1,077	326	10
1.000000	410	1,098	28

'Partition' = 2_Testing	0.000000	1.000000	$null$
0.000000	1,240	342	5
1.000000	405	1,033	26

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.413
1.000000	0.392

'Partition' = 2_Testing	
0.000000	0.371
1.000000	0.448

Confidence Values Report for $KNNP-hasinsurance

'Partition' = 1_Training	
Range	0.5 - 0.955
Mean Correct	0.73
Mean Incorrect	0.633
Always Correct Above	0.909 (2.82% of cases)
Always Incorrect Below	0.5 (0% of cases)
90.56% Accuracy Above	0.727
2.0 Fold Correct Above	0.876 (68.18% of cases)

'Partition' = 2_Testing	
Range	0.5 - 0.955
Mean Correct	0.737
Mean Incorrect	0.634
Always Correct Above	0.909 (3.41% of cases)
Always Incorrect Below	0.5 (0% of cases)
90.78% Accuracy Above	0.727
2.0 Fold Correct Above	0.883 (68.18% of cases)

الشكل رقم 5.9: مخرج من مصنف k-أقرب الجيران

في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

الجدول رقم **1.9**: مقارنة **k**-أقرب الجيران بمصنفات أخرى.

	الدقة (تدريب)	الدقة (اختبار)	حساسية (اختبار)	خصوصية (اختبار)
k-أقرب الجيران	٪73.8	٪74.5	٪70.6	٪78.1
الانحدار اللوجيستي	٪72.19	٪73.1	٪73.6	٪72.3
آلة متجهة الدعم	٪79.25	٪72.93	٪73.6	٪74.4
شبكة محايدة	٪72.74	٪72.89	٪72.60	٪73.2
شجرة تقسيم	٪75.82	٪73.45	٪75.61	٪71.1

مصنف بايز الساذج

يعد مصنف بايز الساذج (Naïve Bayes Classifier) مصنفاً واضحاً وبسيطاً للغاية - على ما يبدو - أثبت نجاحاً ملحوظاً في تطبيقات، مثل عملية مصفاة البريد المزعج، وتصنيف الوثيقة. وظل يستخدم لأكثر من 40 عاماً - وإلى عهد قريب جداً - في معظم تطبيقات استرجاع المعلومات (Lewis 1998). وهو يعمل على الافتراضات غير الواقعية التي تفيد بأن (أ) مساهمة جميع متغيرات المتنبئ في عموم التنبؤ أو التصنيف هي مهمة على نحو متساوٍ، وأن (ب) تأثيرات المتنبئات مستقلة عن بعضها بعضاً. وتسمح هذه الافتراضات غير الواقعية، التي تمنح المتنبئ بايز اسمه، بأن يكون كفوءاً حسابياً، وأن يتطلب بيانات تدريب قليلة جداً، لتطوير تقديرات مَعْلم ما؛ فهو غالباً ما يقوم بأداء جيد، مقارنة بالخوارزميات الأكثر تعقيداً، والكثيفة حسابياً، على الرغم من الافتراضات غير الواقعية التي يستند إليها (Rish, 2001, Zhang 2004).

وفي أي مشكلة تصنيف، ثمة فئة نتيجة، نحاول التنبؤ بها، ومجموعة من متغيرات المُدخل التي نستخدمها لبناء هذا التنبؤ. إننا بصدد القيام بتقدير احتمالية الفئة التي منحت متغيرات المُدخل. ومن ثم، فإن نظرية بايز، تعيد كتابة مسألة التصنيف على النحو التالي:

$$P(Y = y|X = x) = p(Y = y)p\ (X = x|Y = y)/p(X=x)$$

وهذا يعني أن احتمالية النتيجة الممنوحة للمدخل (أو المدخلات)، هي ثمرة احتمالية النتيجة واحتمالية المدخل (أو المدخلات) الذي منح النتيجة، مقسوم على احتمالية المدخلات. وإذا كانت هناك متغيرات متعددة في متجهة X، فسنقوم ببساطة، بمضاعفة الاحتماليات المشروطة. ويمكننا فعل ذلك بالنسبة إلى كُلّ فئة من فئات *Y*، ومن ثم تخصيص لكُلّ حالة لذلك الصنف من أصناف *Y* الذي تعد احتماليته المقدرة (أو «احتماليته الخلفية») الأكثر ارتفاعاً. كما يستخدم مصنف بايز بيانات التدريب لتقدير قيم المَعْلمات على الجانب الأيمن من المعادلة المذكورة أعلاه، ثم تطبيق هذه التقديرات لاختبار البيانات من أجل تصنيفها (Lewis, 1998).

إن بايز الساذج يختلف عن الانحدار من ناحيتين مهمتين:

أولهما: أنه لا يعالج أي واحد من هذه المتنبئات باعتبارها أكثر أهمية من أي متنبئ آخر، والأمر الذي تقوم به - في الأساس - المعاملات في نموذج انحدار لوجيستي، من خلال التصرف كترجيحات، يتم بواسطة كُلّ قيمة متغيرة ما.

ثانيهما: بينما تقدر نماذج الانحدار آثاراً جزئية من المتغيرات - المتوسط المستقل للتأثير الهامشي لكُلّ متغير عندما تبقى قيم المتغيرات الأخرى ثابتة - يسمح بايز الساذج للاحتمالات المشروطة للمتنبئ باستقلاليتها بعضها عن البعض على نحو تام.

مثال في «الرابدماينر» أو المنقّب السريع

لقد كتبت الروتينات من أجل أداء تصنيف بايز الساذج بالنسبة إلى R (وهو تحكم بايز الساذج في حزمة أكبر لـ 1071e) وماتلاب (MATLAB). وهناك تطبيق آخر له في الخادم (Server) إحصائيات نموذج الحزمة الإحصائية للعلوم الاجتماعية، ومختصرات (Macros)، بما أن استعمالها كُتب لأجل نظام التحليل الإحصائي (SAS)، والبيثون (Python).

وفي هذا المثال المعمول به سنقوم ببناء مصنف بايز ساذج باستخدام منقب سريع لحزمة برمجيات مجانية) (بحيث يجري تحميله بسهولة انطلاقاً من الموقع

(http://:rapidminer.com)). ومثله في ذلك مثل منمذج إحصائيات الحزمة الإحصائية للعلوم الاجتماعية الذي نوقش أعلاه، يعمل المنقّب السريع عبر الصياغة السهلة الاستعمال، للتيارات والعقد. ومع ذلك، ينبغي على القارئ ملاحظة أن خوارزمية بايز الساذج، غير مشمولة في نسخة المنقّب السريع، القابل للتحميل بشكل منفصل عن طريق سوق امتدادات المنقّب السريع. وعليه ابحث عن سوق الامتدادات تحت قائمة «ساعد» عند فتح واجهة برنامج المنقّب السريع.

نستخدم البيانات من مسح المجتمع الأميركي للتنبؤ بوضع التأمين الصحي. وتمت معاينة البيانات عشوائياً وموازنتها بحيث تشمل كُلّ من الأفراد المؤمّنين وغير المؤمّنين، 50٪ من الحالات. وكما ناقشنا ذلك سابقاً، إن القيام بعملية موازنة البيانات على مستوى الحصيلة يعد في الغالب فكرة جيدة عند أداء اختبار مصنف ما. وإن القيام بهذا، يزيل من المصنف إغراء سلك السبيل السهل من أجل تقليل معدل الخطأ من خلال تصنيف كُلّ الحالات ببساطة، على أنها تنتمي إلى الصنف المهيمن.

وينبغي اتّخاذ العديد من الخطوات الأولية كي يشتغل مصنف بايز الساذج بطريقة أكثر سلاسة في المنقّب السريع:

أولاً: يعمل الحافر السريع في تجربتنا، بشكل أفضل، وأسرع بكثير، إذا كانت المتنبئات المستمرة المتفردة باستمرار، سابقة لأوانها، على الرغم من أن بايز الساذج يستطيع - نظرياً - أن يتعامل مع متنبئات مستمرة (حساب الاحتمال المشروط من توزيع غاوسي).

ثانياً: يقوم المنقّب السريع بقراءة المتغيرات جميعها بشكل افتراضي بقيم رقمية باعتبارها متغيرات مستمرة. وبتعبير آخر، ينبغي تسجيل المتغيرات الفئوية، والمتغيرات الوهمية من حالة أرقام إلى متغيرات سلسلة (ذات قيم سلسلة) حتى رابدماينر من قراءة هذه المتغيرات بصورة صحيحة.

وبعد تمييز المتغيرات المستمرة، وإنتاج قيم سلسلة، نخصص 70 ٪ من بياناتنا لتدريب النموذج و30٪ لاختباره. وبعد ذلك نقوم بتشغيل نموذج بايز الساذج. وفي المنقّب السريع، يظهر هذا على الشاشة كما هو مبين في الشكلين رقم 6.9 و رقم 7.9.

وتشير مصفوفة الارتباك (الجدول رقم 2.9) إلى أن لدى النموذج دقة شاملة تصل إلى 72.42٪ في بيانات الاختبار، مما يدل على أنه قادر على المنافسة مع المصنفات الأخرى، مثل k- أقرب الجيران التي سبق لنا فحصها.

كما يمنحنا المنقّب السريع أيضاً تقديرات توزيعات احتمالية، في جدول توزيع نموذج (الجدول رقم 3.9). وتعد هذه التقديرات - كما سيذكر ذلك القارئ - احتمال الخاصية لعضوية صنف معين، وليس العكس (لهذا، فإن الاحتمالات لا يصل إلى 100). وينبغي قراءتها على النحو التالي؛ فاحتمال أن تكون حالة ما بيضاء، على اعتبار أنها مؤمّنة، هو 732.، واحتمال أن تكون بيضاء، باعتبارها غير مؤمنة هو، 0.513 ومن ثم، يشكل البيض الأغلبية لدى الأشخاص المؤمّنين وغير المؤمّنين، غير أن تمثيليتهم مفرطة بين الأشخاص المؤمّنين. وعلى النقيض من ذلك، إن احتمال حصول السود على تأمين، هو 0.099، في حين إن احتمال عدم منح السود أي تأمين، هو0.125، مما يشير إلى أن تمثيلية الأميركيين الأفارقة مفرطة بين أولئك الذين يفتقرون إلى تأمين صحي.

وإذا ما اخترنا، فيمكننا التقدير انطلاقاً من جدول التوزيع، احتمالية أن يكون لدى فرد ما مزيجاً معيناً من الخصائص، إما ضمن خانة المؤمّنين أو خانة غير المؤمّنين. خذ، على سبيل المثال، شخصاً أسوداً، أعزب، وحائزاً على درجة الباكالوريوس، ويعيش بالمنطقة الوسطى للولايات المتحدة، وغير إسباني، ومن غير المولودين بالخارج، ويملك منزلاً، وليس مخضرماً، وله عمل، وعمره 27 عاماً، ويعيش في منزل، ويجني 70000 سنوياً. يمكننا ضرب خط أساس احتمال النتيجة في احتمالات هذه الخصائص التي تم الإعلان عنها في الجدول أعلاه، مرتين: مرة يمنح فيها التأمين، ومرة لا يمنح. وفي كلتا الحالتين، تُضرب فيها الاحتمالات أيضاً في الاحتمالات المسبقة للنتيجة (0.50 بالنسبة إلى كُلّ من مسألة وجود التأمين الصحي وعدمه، مع الأخذ بعين الاعتبار توازن البيانات):

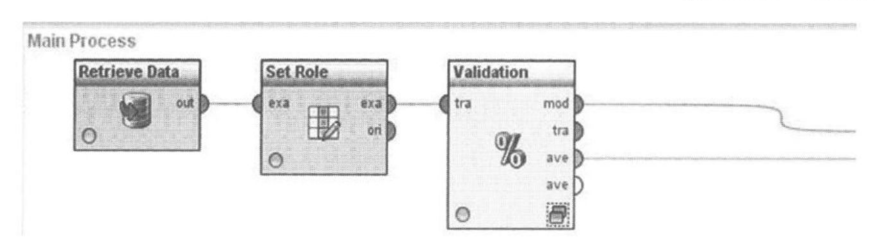
أرجحية احتمال تأمين الفرد = 0.888×0.396×0.144×0.388×0.099×0.50
0.00000298 = 0.172×0.111×0.490×0.911×0.477×0.761×0.962×

أرجحية احتمال تأمين الفرد = 0.709×0.400×0.076×0.492×0.125×0.50
0.000005175 = 0.169×0.295×0.682×0.970×0.532×0.553×0.805×

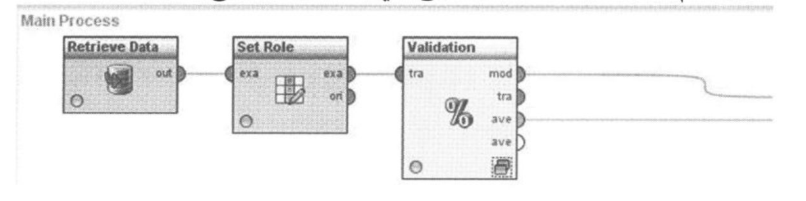
وانطلاقاً من هذا الحساب، يمكن تخمين إمكانية أن يكون هذا النوع من الفرد، غير مؤمن أكثر مما يكون مؤمناً. ولاستجلاء هذه المعلومة أكثر، يمكن تحويل هذه الأرجحيات إلى احتماليات:

احتمالية تأمين الفرد = 0.00000298 /(0.000005175+0.00000298)= ٪35.83 = 0.3583

احتمالية عدم تأمين الفرد = 0.000005175 /(0.000005175 + 0.00000298) = 0.6416 = ٪64.16



الشكل رقم 6.9: بناء تيار بايز ساذج في المنقّب السريع (الإطار الأول).



الشكل رقم 7.9: بناء تيار بايز ساذج في المنقّب السريع (الإطار الثاني).

الجدول رقم 2.9: مصفوفة الارتباك من مصنف بايز الساذج.

	مؤمن حقيقي	غير مؤمن حقيقي	دقة
مؤمن متنبأ	83,239	33,278	٪71.44
غير مؤمن متنبأ	28,256	78,326	٪73.49
مجموع معدل الدقة			٪72.42

الجدول رقم 3.9: جدول توزيع النموذج من مصنف بايز الساذج.

الصفة	معلم (X)	Pr (X)| مؤمّن	Pr (X)| غير مؤمّن
العرق	أبيض	0.732	0.513
العرق	أسود	0.099	0.125
العرق	آسيوي	0.048	0.052
العرق	أميركي أصلي	0.007	0.016
العرق	إسباني	0.112	0.291
العرق	آخر	0.001	0.002
الحالة الاجتماعية	أعزب	0.388	0.492
الحالة الاجتماعية	متزوج، الزوج حاضر	0.449	0.310
الحالة الاجتماعية	متزوج، الزوج حاضر	0.012	0.029
الحالة الاجتماعية	منفصل	0.013	0.030
الحالة الاجتماعية	مطلق	0.080	0.120
الحالة الاجتماعية	أرملة	0.058	0.019
التعليم	أقل من المستوى الثانوي	0.331	0.344
التعليم	المستوى الثانوي	0.209	0.306

التعليم	مستوى كلية ما	0.171	0.203
التعليم	درجة الزميلة	0.059	0.050
التعليم	مستوى الباكالوريوس	0.144	0.076
التعليم	درجة التخرج الجامعي	0.088	0.022
المنطقة	المحيط الهادي	0.154	0.186
المنطقة	الجبال	0.070	0.082
المنطقة	الوسطى	0.396	0.400
المنطقة	شمال شرق	0.187	0.120
المنطقة	المحيط الجنوبي	0.192	0.213
إسباني	لا	0.888	0.709
إسباني	نعم	0.112	0.291
مولود بالخارج	لا	0.962	0.805
مولود بالخارج	نعم	0.038	0.195
صاحب منزل	نعم	0.761	0.553
صاحب منزل	لا	0.239	0.447
الجنوسة	أنثى	0.523	0.468
الجنوسة	ذكر	0.477	0.532
حالة التخضرم	لا	0.911	0.970
حالة التخضرم	نعم	0.089	0.030
الحالة الوظيفية	ليس موظف	0.510	0.318
الحالة الوظيفية	موظف	0.490	0.682
العمر	دون سنّ 10	0.132	0.058
العمر	10–18	0.113	0.070

العمر	30-18	0.111	0.295
العمر	37-30	0.047	0.134
العمر	44-37	0.085	0.124
العمر	50-44	0.085	0.109
العمر	56-50	0.093	0.100
العمر	64-56	0.114	0.093
العمر	+64	0.193	0.016
الدخل الأسري	دون 34,300$	0.240	0.426
الدخل الأسري	50,000-$34,300	0.135	0.190
الدخل الأسري	70,900-$50,000	0.172	0.169
الدخل الأسري	106,000-$70,900	0.205	0.127
الدخل الأسري	أزيد من 106,000$	0.247	0.087

وبعبارة أخرى، من الأرجح تقريباً، أن يكون هذا النوع من الفرد غير مؤمّن بمقدار مرتين أكثر من نسبة كونه مؤمّناً، وأن بايز الساذج سيخصصهما لصنف غير المؤمّن.

وقد رأينا إمكانية أن يكون بايز الساذج مصنفاً كفء، ودقيقاً، ومفيداً. ويقارن بشكل جيد مع الخوارزميات الأكثر تعقيداً، كما أنه يفهم بسهولة أكثر من مصنفات عديدة أخرى، التي تعمل أكثر باعتبارها «صناديق سوداء». وننتقل الآن من إحدى أبسط خوارزميات التصنيف إلى الأكثر تعقيداً: آلة متجهة الدعم.

آلة متجهة الدعم

تعد آلات متجهة الدعم (SVMs) نوع آخر من المصنف. وتم تطوير خوارزمية آلة متجهة الدعم في أوائل التسعينيات من قبل الباحثين في مختبرات بيل (Bell

Laboratories)، فلاديمير فابنيك (Vladimir Vapnik)، (Boser, Guyon, and Vapnik 1992) وتقديمها في شكل عصري في عام 1995 من قبل فابنك وزميله كورينا كورتس (Corrina Cortes) (Cortes, and Vapnik, 1995). وقد جرى تطويره في البداية باعتباره مصنفاً ثنائياً، ومنذ ذلك الحين، تم توسيع إطار آلة متجهة الدعم لتصنيف متعدد الفئات، والانحدار، والتجميع، واكتشاف الشاذ من الحالات، بل وانتقاء سمة نفسه. ومع ذلك، يبقى استعمال التصنيف الثنائي الأكثر شيوعاً. ونركز على هذا التطبيق فيما يلي. أصبحت آلة متجهة الدعم الأجر المعياري في مجالات من قبيل الصورة، وتصنيف النص، وتعرف الحروف، وأثبت أهميته القصوى في العلوم الطبية الحيوية لتصنيف البروتين والكشف عن السرطان. ومع ذلك، لم يظهروا - وبشكل محدود - إلا مؤخراً في العلوم الاجتماعية، في حقول مثل الشؤون المالية (Gavrishchaka, and Banerjee 2006)، والديموغرافيا (Kostaki et al. 2011)، والتسوّق (Cui, and Curry 2005).

ولفهم ما تقوم به آلة متجهة الدعم، يجب أولاً اعتبار مجموعة من النقاط في حيز، المنقسمة إلى فئتين؛ فآلة متجهة الدعم - مثلها مثل مصنفات أخرى - تبحث عن مبدأ، يقسّم هذه النقاط إلى مجموعات بأقل قدر ممكن من الخطأ. وإذا وجدت نقاطنا في حيز ثنائي الأبعاد، فسيكون هذا الفاصل خطاً ما، ويكون مسطحاً في حيز ثلاثي الأبعاد. وفي أبعاد أكبر من هذه، سيكون الفاصل مسطحاً بشكل مفرط للغاية. وبما أن أجهزة الدعم الآلي تبحث دوماً عن مصنف في حيز متعدد الأبعاد، فهي عادة ما تسعى إلى وصف السطح المفرط في الانبساط (أو سطح القرار) التي ستميز بشكل أفضل، بين مجموعتينا. فكر في حيز متعدد الأبعاد مليء بالنقاط الحمراء والزرقاء، حيث الألوان ليست مختلطة تماماً، وإنما وجود مناطق ذات نقاط زرقاء بالأساس، ومناطق أخرى ذات نقاط و«حدود» بالأساس، حيث يفسح لون واحد المجال إلى آخر. إن سطح القرار هو سطح ذو بعد -n (n-dimensional)، قادر على فصل - بقدر الإمكان - مناطق النقاط الزرقاء عن مناطق النقاط الحمراء. والسؤال المطروح، هو أين ينبغي وضع سطح القرار؟

إن أجهزة الدعم الآلي لا تستخدم نقاط البيانات المتاحة جميعها لمعرفة كيفية فصل البيانات، بخلاف تقنيات الانحدار أو العديد من تقنيات التعليم الآلي الأخرى مثل مصنفات بايز أو الشبكات العصبية؛ فهي بدلاً عن ذلك، تستعمل فقط النقاط الأكثر إزعاجاً - النقاط الأقرب إلى «الحدود»، والسطح المفرط في الانبساط الفاصل - للبت في كيفية تشكيل التمييز. وبطبيعة الحال، يتم وصف نقاط البيانات، كُلّ على حدة، بواسطة مجموعة من الإحداثيات، وهي من ثم، متجهات (Vectors). وتعد متجهات كُلّ فئة على حدة، الأقرب في الحيز إلى متجهات الفئة الأخرى، والمستخدمة من قبل آلة متجهة الدعم لإيجاد سطح القرار، تدعى متجهات الدعم (Support Vectors).

ويمكن الآن تقفي أثر عدد لا يحصى من أسطح القرارات أو النسج بين هاتين الحالتين الحاسمتين، ومن ثم، ضرورة اختيار الأفضل منها - أي اختيار أمثلها؛ فأي مبدأ مثالي ينبغي استخدامه، يا ترى؟ يمكننا وصف المسافة بين نقطة وسطر رياضياً. وانطلاقاً من أجهزة الدعم الآلي، نختار سطح القرار ذو المسافة الكبرى بينها وبين متجهات الدعم. تصور مجموعة سطح بين هذه النقاط الحمراء والزرقاء الأقرب إلى الحدّ، والتي تعظّم مسافتها انطلاقاً من تلك المجموعتين من النقاط. وتسمى هذه الفجوة أو المسافة بين متجهات الدعم وسطح القرار، الهامش. وتبحث أجهزة الدعم الآلي عن سطح القرار الذي يعظم الهامش.

إلى حدود الآن، لا تختلف أجهزة الدعم الآلي كُلّ الاختلاف عن باقي الطرق الأخرى المألوفة. وتقوم أجهزة الدعم الآلي بتعقب سطح ما عبر حيز متعدد الأبعاد، الذي يصف بفاعلية، العلاقة بين الخصائص وعضوية المجموعة، وهذا لا يختلف - في واقع الأمر - كُلّ الاختلاف عن الانحدار اللوجيستي، من حيث المبدأ، ولكن تختلف أجهزة الدعم الآلي اختلافاً جوهرياً فقط فيما يتعلق باستخدامها لمجموعات فرعية هامة من الحالات، عوض كلها (ما يجعل أجهزة الدعم الآلي أكثر فاعلية)، ولأن أجهزة الدعم الآلي تعظم المسافة - عوض تقليصها - بين النقاط الرئيسة، وخطّ السطح المفرط في الانبساط الذي تتعقبه.

ولكن افترضنا حتى الآن، أن المعلومات التي بحوزتنا حول حالاتنا أو نقاطنا - أي جمعنا لميزات السمات أو المتغيرات - سيسمح لنا برسم خطّ أو السطح المفرط

في الانبساط عبر النقاط، التي تفصلها إلى مجموعتين متميزتين؛ أي أننا افترضنا أن مجموعاتنا قابلة للفصل خطياً.

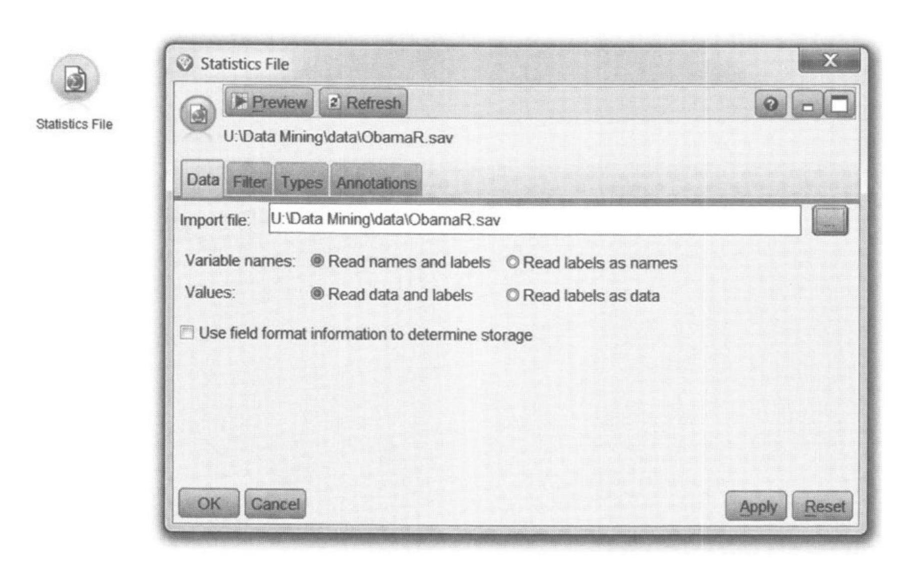
ولكن لا يكون الأمر على النحو في كثير من أحيان. قد تكون لدينا - على سبيل المثال - حالات تكون فيها فئة واحدة (أو لون نقاط) محاطة بالكامل بتلك الفئات الموجودة في فئة مختلفة. وإن صح ذلك فلن يوجد خطّ أي فاصل خطي، يمكن تصوره، والذي قد يسمح بفصل الحالات إلى مجموعاته الخاصة. وما دمنا مقتصرين على الحيز البعدي -n المحددة بمُدخلاتنا (حيز المدخل، في لغة آلة متجهة الدعم)، فإن أي مصنف سيعجز عن تصنيف نسبة كبيرة من الحالات بشكل صحيح.

هنا تظهر جدّة أجهزة الدعم الآلي، إذ يفسر صعوبة هذا التصنيف بالبُعدية المقيدة. وإذا أمكن لنا إسقاط بياناتنا داخل فضاء ذي بعد عالي، فسيكون بإمكاننا فصل هذه الحالات. ويشار إلى هذا الحيز ذي البعد العالي (أو حتى الحيز ذي البعد اللا متناهي)، حيث الحالات منفصلة خطياً، بحيز السمة. كما أن رسم حيز المدخل حيز السمة هو مجرد مسألة تطبيق وظيفة رياضية على البيانات لتحويلها بشكل مناسب إلى حيز ذي بعد عالي.

وتكمن الصعوبة في كون أن خصائص حيز هذه السمة، غير معروفة لدينا، ولهذا فمن غير الممكن عموماً، معرفة الوظيفة الرياضية الحقيقية التي نحتاجها. ولكن الظاهر أن هذا لا يهم في واقع الأمر. وكل ما علينا القيام به، هو تعريف دالة النواة (Kernel Function) التي ستقاربها (التي يشير إليها مطورو أجهزة الدعم الآلي بتعبير خدعة النواة (Kernel Trick)). وثمة العديد من وظائف النواة، وعموماً ستوفر برامج آلات متجهة الدعم المستخدم بعدد قليل من الخيارات، حول النواة الممكن استخدامها. وإن أفضل نواة مؤهلة لهذه المهمة، ليست شيئاً يمكن معرفته عادة في وقت سابق لأوانه) اللهم إلا إذا كنت تجيد رسم البيانات داخل أبعاد عليا في ذهنك)، ولهذا يمكن أن يتم الاختيار فقط عبر التجربة والخطأ.

هنالك صعوبة واحدة بخصوص هذه الخدعة، وهو أنه من المحتمل أن تفوق تناسبيتها البيانات. وبتعبير آخر، قد يؤدي إسقاط البيانات إلى حيز ذي بعد عالي إلى

انفصال خطي كثير للبيانات الخاصة التي بين أيدينا، ولكنها حققت الانفصال عن طريق رفع كوكبة من متجهات دعم، وقد تكون هذه الكوكبة خاصة بتلك البيانات المميزة. وبالتالي فقد أضحى من الضروري أداء الصلاحية المتبادلة عندما يتم توظيف أجهزة الدعم الآلي سواء عن طريق الإبقاء على جزء من البيانات لاختبار نموذج آلة متجهة الدعم، أو عن طريق الصلاحية المتبادلة لطية k-. وسيطلعك هذا عما إذا كان نموذج آلة متجهة الدعم يعمل عند تطبيقه على بيانات أخرى – أي ما إن كان تعميمه أمراً ممكناً.



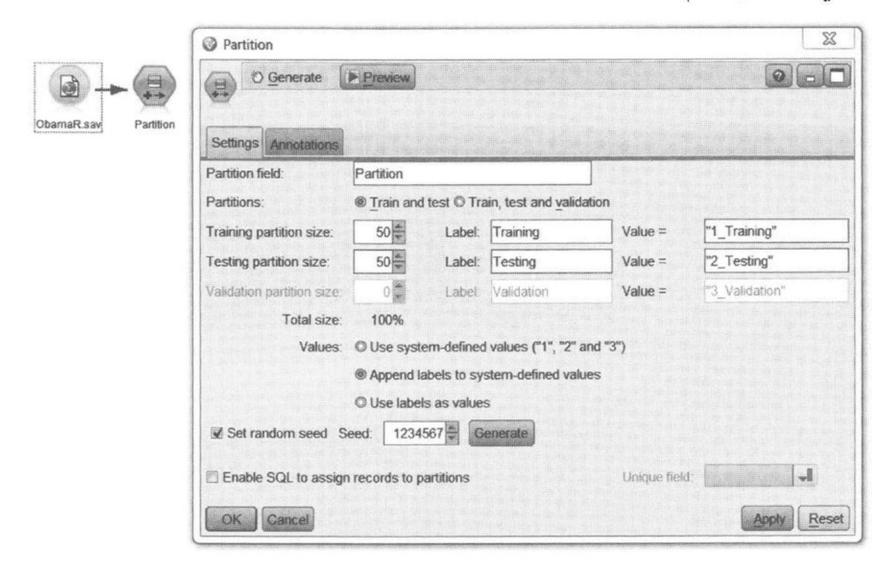
الشكل رقم 8.9: تحميل البيانات لأجل تحليل آلي لمتجهة دعم في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

آلات دعم المتجه في منمذِج الحزمة الإحصائية للعلوم الاجتماعية

لأن أجهزة الدعم الآلي لم تكن موجودة طوال هذا الوقت، ولأن استعمالها لا يزال مقتصراً على مهام متخصصة، لم يتم دمجها في معظم الحزمات الإحصائية التجارية. إن برامج آلات متجهة الدعم متوافرة في R و«الماتلاب» وكذا في عدد من أجنحة التنقيب في البيانات. وإن الحزمة الإحصائية للعلوم الإنسانية قد تم ضمها

أيضاً إلى برنامج منمذجها من برامج التنقيب في البيانات، التي سنقدم توضيحاً بشأنها أدناه.

ولتشغيل آلة متجهة الدعم في المنمذج، نحتاج أولًا إلى انتقاء بعض البيانات. والمنمذج قادر على قراءة عدد من أنواع مختلفة من ملفات البيانات مثل ملفات إكسيل أو ملفات النص؛ فبياناتنا موجودة سلفاً في ملف (.sav) لحزمة الإحصائية للعلوم الاجتماعية SPSS، لذا ننتقي جدول الموارد للوحة العقد وننتقي ملف الإحصائيات. وبعد ذلك تفتح العقدة على الشاشة، التي ننقرها مرتين لانتقاء ملف البيانات الذي نريده. (لتصفح الملف على حاسوبك، اضغط على الزر الأزرق ذي النقاط الثلاث، على يمين علبة نصّ الملف استيراد الملف (Import File) كما هو مبين في الشكل رقم 8.9).



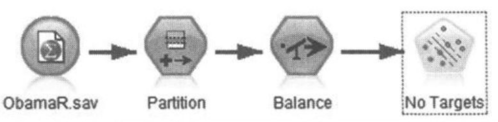
الشكل رقم 9.9: تقسيم البيانات قبل تحليل آلة دعم المتجهة في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

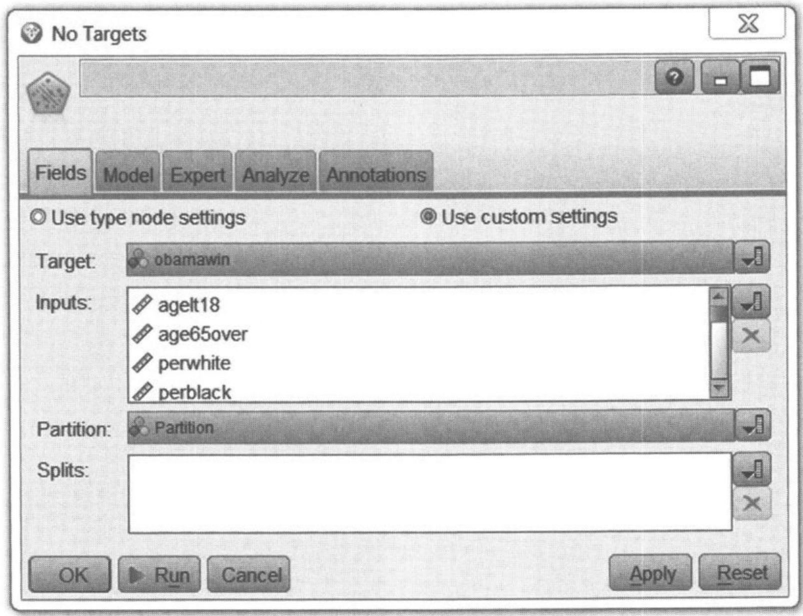
نختار مجموعة بيانات انتخابات 2012 على مستوى المحافظة، التي قمنا بتعديلها لتشمل المتغير الوهمي، المشفر 1 إن عادت 50٪ أو أكثر من أصوات

المقاطعة إلى أوباما و0 فيما عدا ذلك. وعلى العموم، فبياناتنا صيغت مسبقاً في الجزء الأكبر منها. وهذا ليس أمراً ضرورياً بما أن المنمذج يمكنّك من إنتاج متغيرات جديدة أو تحويل تلك الموجودة. كما يسمح لك أيضاً معاينة حالاتها أو إقصاء مجموعة فرعية منها، ولكن تهيئ البيانات هو - إلى حدّ ما - أمر أسهل وأكثر بساطة في برنامج إحصائي معياري مثل الحزمة الإحصائية للعلوم الاجتماعية (SPSS) أو «الستاتا» (Stata)، خاصة إذا كنت مطالباً بالقيام بالعديد من التغييرات، ولهذا نقترح إعداد بياناتك أولاً قبل تحميلها في المنمذج. ومع ذلك، ستتحقق من أن كُلّ المتغيرات مشفرة بشكل صحيح باعتبارها متغيرات مستمر وفئوية، وهكذا. ويمكن القيام بهذا يدوياً أو آلياً من خلال نقر اقرأ القيم (Read Values) في جدول الأنواع في نافذة العقدة لملف الإحصائيات.

وبعدها نقسم البيانات إلى قسمين. ولا بُدَّ من القيام بهذا التقسيم في عقدة مستقلة، بدلاً من انتقائه كخيار ضمن نافذة الإجراء كما في الغامب (JMP). وفي لوحة العقد، قم بانتقاء جدول مجال العمليات، وبعدها انقر تقسيم (Partition). وهذا يستدعي نافذة التقسيم المعروضة في الشكل رقم 9.9. كما يمكّنك المنمذج من إنتاج أجزاء التدريب، والاختبار، والصلاحية أو فقط الجزأين الأولين واختيار جزء البيانات المراد إدراجها في كُلّ واحد، كما سنقتصر على إدراج جزئي التدريب والاختبار بما أن بياناتنا تملك فقط 3.114 حالة.

ومن أصل 3.114 محافظة من محافظاتنا، صوتت حوالي 25٪ لصالحه، فيما صوتت 75٪ ضده (لكن 25 في المائة من المحافظات التي ربحها أوباما، كانت في معظمها محافظات مليئة بالسكّان). ويجد العديد من المصنفات صعوبة مع البيانات غير المتوازنة من حيث النتيجة التي تميل إلى تقليص معدل الخطر من خلال تخصيص كُلّ الحالات أو معظمها للأغلبية. ولمعرفة ما إن كان هذا الأمر مهما بالنسبة إلى أجهزة الدعم الآلي، سنقوم بتحليلات تهم البيانات المتوازنة وغير المتوازنة. وإن موازنة البيانات بسيطة - نوعاً ما - في المنمذج. يكفي نقر جدولة سجل العمليات (Record Ops Tab)، واختيار التوازن (Balance).

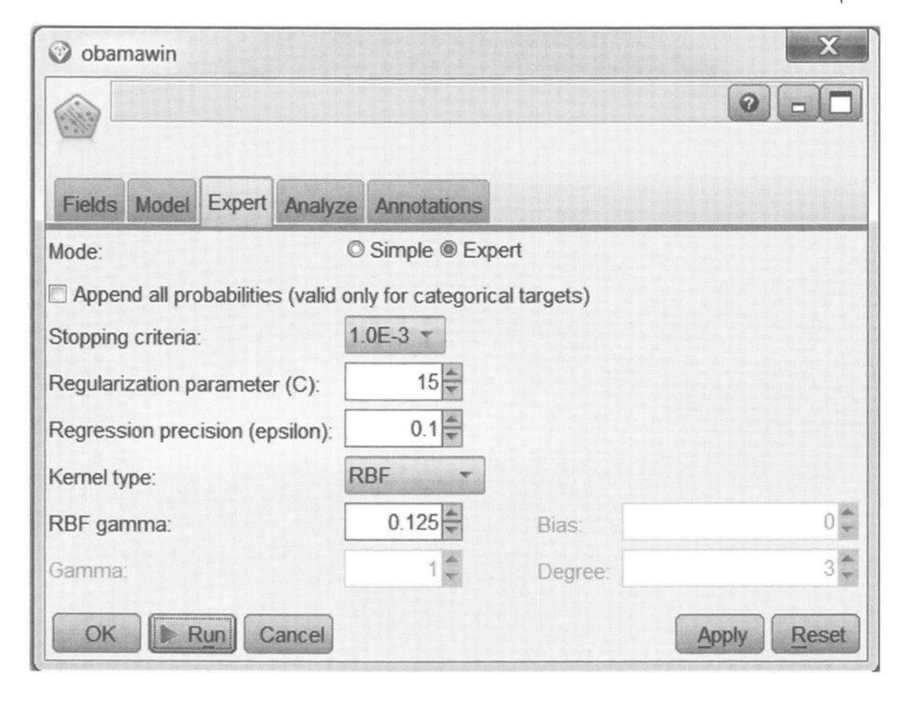




الشكل رقم 10.9: بناء تيار تحليل آلة متجهة الدعم في منمذج الحزمة الإحصائية للعلوم الاجتماعية.

ويتم موازنة المنمذج من خلال تقليص فئة معينة من الحالات (عن طريق معاينة حالات بشكل عشوائي) أو من خلال زيادة فئة أخرى (عبر مضاعفة الحالات). ستحتاج إلى اختيار صيغة تخبر المنمذج عن الحالات التي تريد أن تغير (مثل «فوز أوباما = 1»)، ومن ثم قاسماً مشتركاً يتم من خلاله ضرب الحالات لتحقيق عددك المرغوب فيه. ونضرب فوز أوباما = 0 في 0.4، وفوز أوباما = 1 من الحالات في 1.54 لموازنة البيانات (إلى حدّ ما) على مستوى الحصيلة. ويمكنك هنا اختيار سواء ما إن كنت تريد موازنة البيانات فقط في مجموعة التدريب، أو أيضاً في مجموعتي الاختبار والصلاحية. وإن معاينة البيانات في مجموعة التدريب أمر مفيد فقط إن أردت استعمال مجموعة الاختبار لتوليد نتائج الميول (الذي يمكن للمنمذج توليده بسهولة). وبما أننا لا نريد القيام بذلك، فسنوازن على مستوى المجموعتين معاً.

إننا الآن على استعداد لتشغيل آلة متجهة الدعم، وهذا موجود تحت قائمة النمذجة باعتباره آلة متجهة الدعم. وبعد إضافته إلى التيار، ستحتاج إلى بناء نموذج خاص بك (انظر الشكل رقم 10.9). واشرع بداية في تحديد المتغير الهدف، وبعدها اختيار المتنبئات التي تريد في النموذج، ومتغير التقسيم (المولّد تلقائياً إن كنت قد أنتجت عقدة التقسيم). ثم، قم باختيار خاصيات آلة متجهة الدعم الذي تريد تشغيلها، وذلك عبر اختيار جدولة الخبير في نافذة آلة متجهة الدعم. كما يمكنك فتح مفاتيح الخيار عبر ضبط الوضعية للخبير ضمن هذه الجدولة (انظر الشكل رقم 11.9). وسيمكنك هذا من اختيار نوع النواة ومعلم «غاما» (إن كانت لديك نواة لا خطية)، ومعلم الضبط لـ C وضبط الدقة وقاعدة الإيقاف.

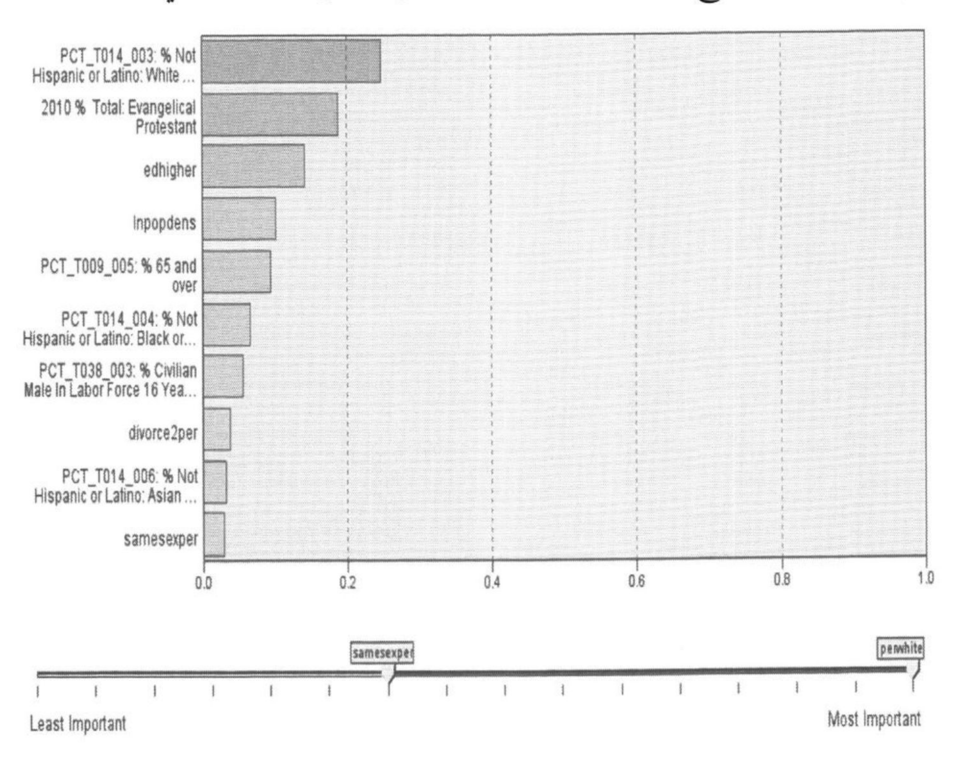


الشكل رقم 11.9: إعداد مَعْلمات آلة متجهة الدعم في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

إن قاعدة الإيقاف تخبر المنمذج عندما تريدها أن تبت في كون التقاء نموذجك. وقد تم تخفيض الافتراض - إلى حدّ ما - ولكن إذا أردت أن تتيقن من بلغوك الحدّ

الأدنى العام، يمكنك إدخال عدد أقل انخفاضاً. وإذا أردت التقاء أسرع، فارفع من قيمة الإيقاف. ولكن المنمذج يعمل - إلى حدّ ما - بسرعة مع بيانات من هذا الحجم، لذا ننصح بملازمة الافتراض.

إن المنمذج يسمح لك باستعمال أربع نوى مختلفة - دالة القاعدة الشعاعية (RBF)، والدالة المتعدد الحدود، والدالة السينية، والدالة الخطية. وإن النوى الخطية لا تسقط البيانات إلى حيز عالي الأبعاد؛ فإذا تناسب ذلك جيداً، فسيكون لديك فقط بيانات لا تحتاج إلى أن ترسم إلى حيز سمة ليتم تصنيفها. أما النوى الأخرى، فلديها كلها مواطن قوتها، ونقترح أن تجرب كُلّ واحدة على حدة، بالإضافة إلى قيم مختلفة من معالمات النموذج، قصد الحصول على أفضل فاصل دون إفراط في التناسب.



الشكل رقم 12.9: أهمية المتنبئ في آلة متجهة الدعم في منمذِج الحزمة الإحصائية للعلوم الاجتماعية. الهدف: فوز أوباما.

وبمجرد اختيارك نواة ما، عندئذٍ يكون الوقت قد حان لضبط مَعْلمات النموذج،

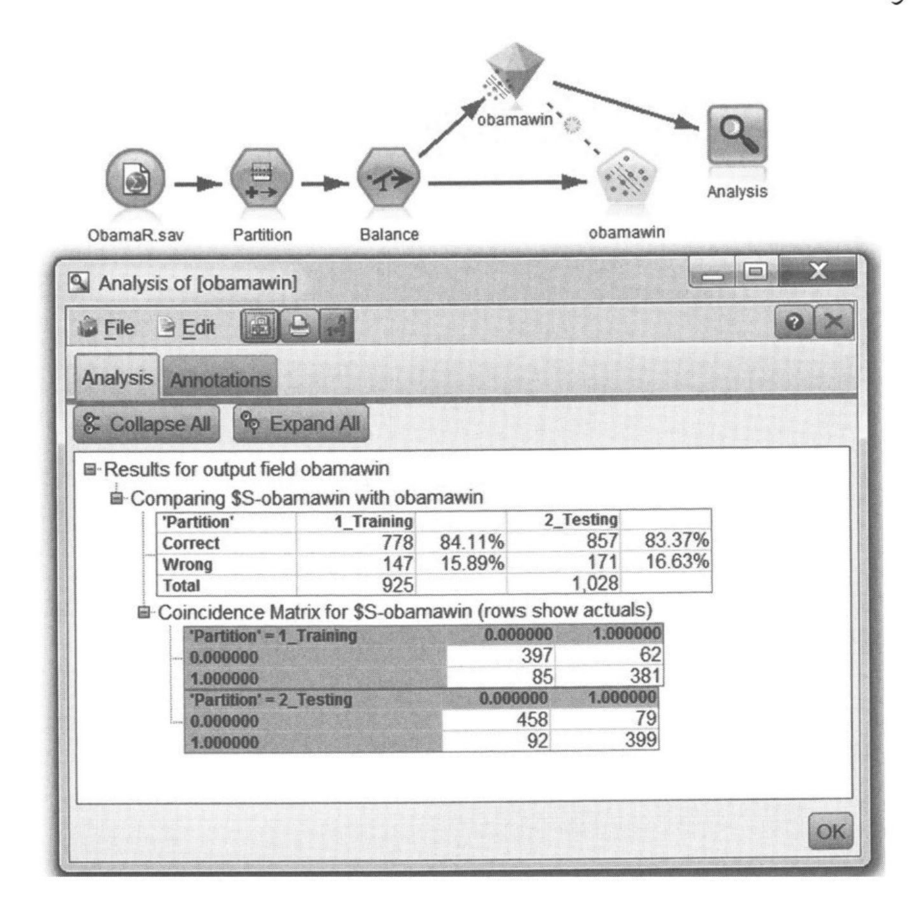
ويكون بإمكانك وضع معلم الضبط لـC ، ودقة الانحدار (إيبسيلون) (Epsilon) بالنسبة إلى أي نوع من أنواع النواة الأربع. وأما دقة الانحدار، فلا تهم إلا نمذجة نتائج مستمرة (حيث المنمذج يؤدي انحدار آلة متجهة الدعم)؛ فهي تخبر النموذج عن حجم الخطأ المقبول، في حين يعد معلم الضبط لـ C، استتباعياً بالنسبة إلى المقايضة بين الدقة في بيانات التدريب والإفراط في التناسب. وتنجم عن القيم العليا لـ تصنيفاً C أكثر دقة، ولكن بإمكانها تقليص قدرة النموذج على التعميم لتشمل بيانات الاختبار. أما بالنسبة إلى دالة القاعدة الشعاعية، والدالة متعددة الحدود، والدالة السينية، فهناك أيضاً مَعْلم «غاما». ومثلها في ذلك مثل C، تنجم عن القيم العليا مزيداً من الدقة على حساب إفراط تناسبي مفترض. وإذا اخترنا النواة المتعددة الحدود، فيمكننا ضبط درجة النواة المتعددة الحدود (الفرضية هي 3). وأخيراً، يمكن للشخص وضع معلم متحيز، مماثل لمتغير ثابت في الانحدار بالنسبة إلى كُلّ من النوى المتعددة الحدود، والنوى السينية.

أي من هذه الإعدادات يجب وضعها؟ يقدم البرنامج بعض الإرشادات، ولكن ضمن هذه المَعْلمات، سيكون من الصعب الإدلاء برأي قبل الحدث. ويمكن للباحث فقط أن يجرب مع إعدادات مختلفة ويختار الإعدادات الأمثل.

قم بتشغيل برنامجك عبر نقر «تشغيل». تظهر «كتلة صلبة» وهي العقدة التي تحتوي نتائج النموذج الذي قمت بتشغيله. ولسوء الحظ، لا يتوافر جزء كامل من حيث المخرج انطلاقاً من آلة متجهة دعم في المنمذج. وإذا انتقيت «احسب أهمية المتنبئ» (Calculate Predictor Importance) في «جدولة حلل» (Analyze Tab) داخل نافذة آلة متجهة الدعم، فسيتم عرض المتغيرات المختلفة مساهمات في الفاصل (الشكل رقم 12.9).

وما نلاحظه هنا هو أن المتغير الأهم في تنبؤ أصوات أوباما، يتمثل في نسبة المحافظات من البيض غير الإسبانيين. وبعد هذا، وفي انخفاض للأهمية بشكل سريع، نجد نسبة البروتستانت الإنجيليين، المتناسبة مع درجة باكالوريوس ما، أو درجة أعلى منها، والكثافة السكانية، ونسبة 65 أو أكبر. ولا يعد المنمذج موثقاً جيداً على نحو خاص، ولسوء الحظ، عندما يتعلق الأمر بوصف مدلول إحصائياته المولدة

(مثل أهمية المتنبئ) على وجه الدقة، أو كيف يتم حسابها، ولكن المعاني حدسية نوعاً ما.



الشكل رقم 13.9: مخرج آلة متجهة الدعم في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

للبحث عن إحصائيات التناسب، اختر المُخرج، وانتقي «عقدة حلل» (انظر الشكل رقم 13.9). وضمن هذه العقدة، في «جدولة حلل»، انقر المربع الخاص «بمصفوفات المصادفة». وسيمنحك هذا، النسب المئوية مصنفة بشكل صحيح في مجموعات بيانات التدريب والاختبار، وكذا مصفوفات الارتباك بالنسبة إلى الذين يمكنك استخدامها معاً لحساب حساسية النموذج، وخصوصيته.

لقد قمنا بتشغيل النماذج على بيانات انتخابات 2012، مستخدمين نوى مختلفة (مغيرين المَعْلمات من أجل أداء أمثل) ومن أجل أن تقوم المقارنة بعملية تشغيل النماذج نفسها، باستخدام الانحدار اللوجيستي، وأشجار التقسيم (أشجار الانحدار والتصنيف في المنمذج)، والشبكات العصبية. ولأجل مقارنة أكثر، نقوم بتشغيل النماذج، مستعملين بيانات متوازنة وغير متوازنة، والنتائج معروضة في الجدولين، رقم 4.9 و5.9.

أما في البيانات غير المتوازنة، فتتفوق آلات متجهة الدعم بقليل من حيث الأداء على المصنفات الثلاثة الأخرى في بيانات التدريب، لكنها أقل أفضلية بشكل واضح في بيانات الاختبار. ولكن هذا لا يعني بالضرورة أن تكون النماذج مفرطة في التناسب. وفي نهاية المطاف، حققوا الأفضلية في مجموعة التدريب لكنها منافسة، وفي أغلب الأحيان أحسن في مجموعة الاختبار أيضاً. وفي هذه البيانات، فقط دوال النواة المتعددة الحدود هي التي تبدو مفرطة في التناسب. أما نواة دالة القاعدة الشعاعية، فتتفوق على كُلّ المصنفات المنافسة في بيانات الاختبار (على الرغم من أن ذلك لا يتم على نحو كبير).

الجدول رقم 4.9: مقارنة أداء آلة متجهة الدعم مع المصنفات الأخرى باستخدام بيانات غير متوازنة.

	الدقة (التدريب)	الدقة (الاختبار)	الحساسية (التدريب)	الحساسية (الاختبار)	الخصوصية (الاختبار)
نواة دالة القاعدة الشعاعية	٪89.8	٪88.9	٪66.45	٪60.72	٪96.03
آلة متجهة الدعم (نواة الدالة المتعدد الحدود)	٪97.03	٪85.95	٪89.67	٪67.06	٪90.78

آلة متجهة الدعم الخطية	87.82٪	88.14٪	60.64٪	58.06٪	95.73٪
الانحدار اللوجيستي	88.3٪	88.02٪	62.5٪	59.21٪	95.27٪
شجرة التقسيم	86.73٪	84.98٪	65.80٪	59.51٪	91.39٪
الشبكة العصبية	87.62٪	88.32٪	59.35٪	58.30٪	95.88٪

الجدول رقم 9-5: مقارنة أداء آلة متجهة الدعم مع المصنفات الأخرى باستخدام بيانات متوازنة.

	الدقة (التدريب)	الدقة (الاختبار)	الحساسية (التدريب)	الحساسية (الاختبار)	الخصوصية (الاختبار)
نواة دالة القاعدة الشعاعية	83.84٪	84.11٪	80.77٪	82.12٪	85.97٪
آلة متجهة الدعم (نواة الدالة المتعدد الحدود)	88.52٪	81.54٪	89.38٪	77.64٪	85.41٪
آلة متجهة الدعم الخطية	84.21٪	81.57٪	84.64٪	82.06٪	81.08٪
الانحدار اللوجيستي	83.32٪	83.25٪	81.87٪	80.91٪	85.71٪
شجرة التقسيم	80.6٪	77.76٪	90.81٪	84.53٪	71.55٪
الشبكة العصبية	85.16٪	82.40٪	88.84٪	86.33٪	78.63٪

وتتألف البيانات - بشكل غير متناسب - من نتائج سلبية، وفي هذه الحالات تميل المصنفات في الغالب، إلى الفشل عبر سوء تصنيف الإيجابيات. إذن، إن قياس حساسية النموذج (نسبة الإيجابيات المصنفة بشكل صحيح) مهم للفحص. ومن المهم خاصة، فحص الحساسية في مجموعة الاختبار. وهنا تبين آلات متجهة الدعم عن أفضليتها. وباستثناء آلة متجهة الدعم الخطية، التي لا تستغل رسم البيانات في حيز عالي الأبعاد، تبقى آلات متجهة الدعم أفضل من الطرق الأخرى في العثور على النجاحات، وإن آلات متجهة الدعم المفرطة في التناسب، موفقة فيها خاصة. إذن، لا تستخدم آلات متجهة الدعم كُلّ البيانات، وإنما فقط حالات محدودة، بغية تحقيق تصنيف أمثل. ونتيجة لذلك، فإنها أقل عرضة للخطأ من خلال تخصيص معظم الحالات إلى الفئة الغالبة.

وعندما نتحول إلى البيانات المتوازنة، نجد أن آلة متجهة الدعم إلى جانب نواة دالة القاعدة الشعاعية أفضل - نوعاً ما - من الانحدار اللوجيستي، والشبكات العصبية من حيث الدقة العامة. ومرة أخرى، تبدو دوال آلة متجهة الدعم موسومة بالإفراط في التناسب قليلاً. وإن آلة متجهة الدعم لدالة قاعدة الشعاع، تفوق بقليل مصنفات المنافس من ناحية الحساسية، ولكن شجرة التقسيم والشبكة العصبية يفوقانها من حيث الحساسية. وأما مسألة عدم تفوق آلات متجهة الدعم في الأداء - بشكل كبير - على منافسيها، فقد تكون تلك دالة بيانات (يمكن أن تكون قابلة للانفصال خطياً مع عدم منح آلات متجهة الدعم أية امتياز) أو بيانات تنفيذ خاص في المنمذج (وهذا ليس شيئاً مرناً خاصة، من ناحية تعديل المتغير). ومن ناحية أخرى، إن خوارزميات المنافس، جيدة جداً في تصنيف البيانات في العديد من الظروف.

إننا نقترح أن يجرب الباحثون آلة متجهة الدعم في أوساط البحث في العلوم الاجتماعية، واستعمالها في حالات تتفوق فيها على المصنفات. وبالإمكان استخدام آلات متجهة الدعم لتوليد درجات الميل مثلاً. وقد يكون هذا مفيداً لغايتك، وقد لا يكون كذلك. وإن تطبيق منمذج آلات متجهة الدعم ليست مفيدة خاصة من ناحية تزويدها لنا بمعلومات عن علاقة السمات بالنتيجة، ذلك بأنها لا تخبرنا عن النموذج الذي تبنيه من أجل التصنيف. ولكن على العموم، إن قوة آلات متجهة الدعم الحقيقية

- أي قدرتها على رسم البيانات في حيز عالي الأبعاد، عبر دالة نواة - يجعلها مبهمة تماماً. وفي هذا الصدد، فهي تشبه الشبكات العصبية؛ وإن تحويل النواة ليس في الواقع علبة سوداء ولا هو شفاف. ومع ذلك، ينبغي اسكتشاف فائدة آلات متجهة الدعم بما أنها أثبتت نفسها بشكل كبير، على أنها بارعة في تصنيف البيانات المعقدة في العديد من الأوساط العملية.

أمثلة التنبؤ عبر مصنفات متنوعة

لقد راجعنا عدداً من خوارزميات التصنيف وتم تطوير طرق أخرى عديدة (اثنتين منها - أشجار التقسيم والشبكات العصبية - سيتم تغطيتهما بالتفصيل في الفصول اللاحقة). وأكثر من ذلك، أن بعض الباحثين، قد طوروا العديد من المتغيرات في كُلّ طريقة على حدة. وإن الأسئلة الطبيعية التي تطرح في هذه المرحلة هي: أيهما أفضل؟ وأي متغير ينبغي استخدامه؟ هل هناك متغير أكثر دقة وخال من الغموض؟ وهل تعتمد على البيانات؟ وإن صح ذلك، فهل هناك قواعد صعبة وسريعة (أو حتى قواعد بديهية) لاختيار المصنف إن كنت أعرف شيئاً عن البيانات؟

لسوء الحظ، جواب كُلّ هذه الأسئلة المطروحة هو: النفي على الإطلاق؛ أو بالأحرى، هذا يتوقف على سياق الحال، لكن ليس على مميزات البيانات في حدّ ذاتها. فقد نرى - عوضاً عن ذلك - أنه من الأفضل تصور الاختيار باعتباره مسألة عملية. علاوة على ذلك، إن المصنف الأفضل - في بعض الأحيان - هو مسألة تتعلق «بالأشياء الأخرى» التي يقوم بها المصنف أثناء عملية التصنيف. سنقضي، على سبيل المثال، بعض الوقت في أشجار التقسيم، ليس لكونها قوية بالضرورة في مهام التصنيف (على الرغم من أنهم في الغالب كذلك) أو لأن نتائجها تُعمَّم جيداً على العينات الخارجية (على الرغم من أنها كذلك)، لكن لأن نتائجها مرنة للغاية. وتسمح لنا دراسة بنية الشجرة ذاتها، التعرُّف - نوعاً ما - إلى العلاقة بين النتيجة ومقاييس المتنبئ، كما تطلعنا على أشياء عنها، لا يطلعنا عنها - بالضرورة - الانحدار اللوجيستي. ويمكن أيضاً استعمال الشبكات العصبية لمساعدتنا على فهم العلاقات المعقدة اللا خطية، خصوصاً لما يتم دمجها مع برمجيات التصور مثل «الغامب برو». وعند هذه النقطة، لا يزودنا كُلّ من مصنفات k-أقرب الجيران، وآلات متجهة الدعم

بكل هذه المعلومات المتعلقة بالمتنبئات نفسها على الرغم من أنها (جدلاً) أفضل في بعض مهمات التصنيف المعقدة.

ولكن ثمة جواب آخر لهذا السؤال، يتمثل في عدم ضرورة اختيار طريقة واحدة أفضل. وممكن جداً مزج نتائج العديد من خوارزميات التصنيف في نتيجة واحدة نهائية. ويبقى المشكل في القيام بهذه بطريقة تتوسل بنقاط القوة النسبية، بدلاً من نقاط الضعف النسبية لمختلف تقنيات التصنيف (Xu, Suen, Kryznak 1992). وقد يستطيع الشخص الحصول على تصنيف ما، أكثر قوة من أي تقنية بمفردها، وذلك من خلال مزج الطرق. ولكن هذه النتيجة ليست مضمونة، وربما لا يعد هذا حتى الأساس المنطقي الأفضل لمزج المصنف. وثمة أساس منطقي ثانٍ، يساعد على مزج المصنفات على تقليص احتمال أن تؤدي النتائج التمييزية انطلاقاً من اختلافات أي طريقة من طرق التصنيف، إلى القرار النهائي للتصنيف. ومن الأرجح أن يعطي مزج المصنفات نتيجة أكثر سلاسة (Smoother)، نتيجة قد تكون قادرة على التعميم بشكل أفضل نوعاً ما. وبهذا المعنى، تصبح عملية مزج المصنفات شبيهة أكثر بالتعبئة (أو بغابات عشوائية) مقارنة بالتعزيز بحسب منطقها.

عملية مزج المصنفات في منمذِج الحزمة الإحصائية للعلوم الاجتماعية

في منمذِج الحزمة الإحصائية للعلوم الاجتماعية، تتحقق عملية مزج المصنفات بسهولة عبر استعمال عقدة المصنف الذاتي. وهذه عقدة مستقلة، تسمح للمستعمل بانتقاء مصنفات مختلفة والمَعْلمات التي تتحكم في كيفية اختيارها ومزجها. وباستعمال هذه العقدة، سيكون الإغراء - في الغالب - متجها ببساطة نحو استعمال الإعدادات الافتراضية، الخاصة بكل نوع من المصنفات المستخدمة؛ وإذ نُحذِر من هذا، فإننا نقترح بدلاً من ذلك، الضبط بعناية كُلّ نموذج على حدة ليصبحوا نماذج مثالية قبل التصنيف. وإن استعمال أدوات التنقيب في البيانات - كما هو الحال دائماً - بعناية وبحكمة، هو أمر مفضل.

ومرة أخرى نبدأ بانتقاء بياناتنا من بيانات مسح المجتمع الأميركي التي سنستعملها في توقع تغطية التأمين الصحي. ونتيقن - في جدولة النوع (لوحة عمليات

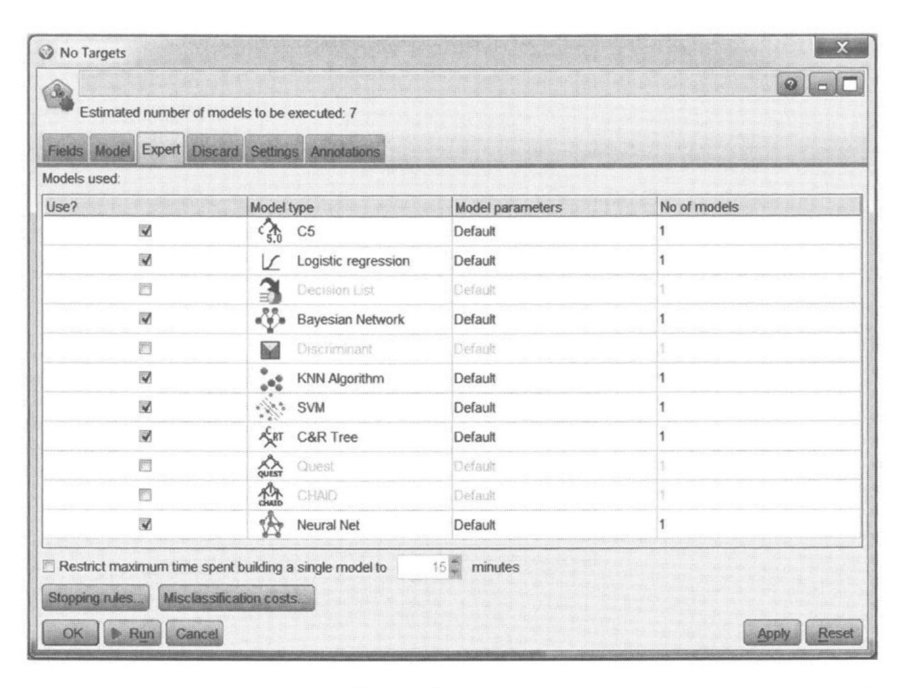
المجال) - من أن يخصص للمتغيرات، المستوى الصحيح للقياس، وبعدها نقسم البيانات 50٪: 50٪ في مجموعتي التدريب والاختبار (جدولة التقسيم، ولوحة عمليات المجال). وبعد ذلك، وفي لوحة النمذجة، نختار المصنف الذاتي.

وبعد انتقاء النتائج، والمتنبئ ومتغيرات التقسيم، نستمر في تأسيس مصنفاتنا. ونقترح بدلاً من التوجه أولاً إلى جدولة النموذج، نقر «خبير»، والمضي ناحية اليمين لتحديد المصنفات وإعداداتهم. وإن القيام بذلك، يقدم لك شاشة كتلك المعروضة في الشكل رقم 14.9. ويختار لك البرنامج تلقائياً، مصنفاً من أصل ثمانٍ مصنفات، بحيث يكون لكُلّ واحد إعداداته الافتراضية. وستحتاج إلى تحرير هذه الشاشة لضم النماذج التي تريد. ومن المهم ملاحظة بضعة أشياء ينبغي أن توجه هذا القرار.

أولاً: يمكنك الحصول على نسخ متعددة لمصنف واحد، كُلّ بإعدادات معلم مختلفة. هل تعذر عليك البتّ في ما إذا كنت تريد آلة متجهة الدعم ذات دالة قاعدة شعاعية أو نواة سينية؟ لا بأس - قم بضم كُلّ واحدة منها. وبإمكانك البتّ فيما إذا كنت تريد ضم كُلّ النتائج أو فقط الأفضل منها، في تنبؤك النهائي.

ثانياً: تذكر أن مزيداً من النماذج، يعني معالجة بيانات أكثر. وهذا يعني - بدوره - بكل تأكيد، مزيداً من الوقت، كما يعني أيضاً - ولسوء الحظ - احتمالية أكبر لتجميد البرنامج أو انهياره. ويستحسن القيام بتجربة بسيطة حول هذا قبل تجريب كُلّ شيء دفعة واحدة. وهذه أيضاً فائدة من فوائد خوض التجربة مسبقاً مع نماذج فردية بشكل عرضي.

سنقوم بانتقاء خمس نماذج - الانحدار اللوجيستي، وk-أقرب الجيران، وآلة متجهة الدعم وشجرة تقسيم واحدة (وتسمى هنا شجرة C&R) - وبتعديل إعدادات كُلّ واحد منها. وسنعود الآن إلى جدولة النموذج (الشكل رقم 15.9) والقيام بتعديل القواعد لمزج النتائج.

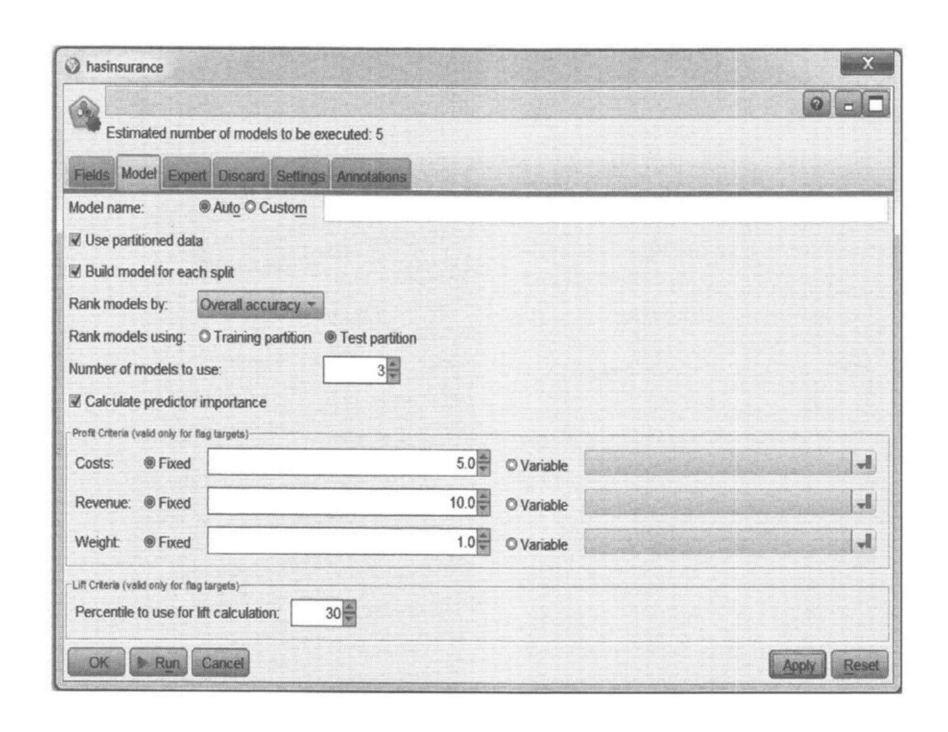


الشكل رقم 14.9: انتقاء المصنفات من أجل الأمثلة، باستخدام عقدة المصنف الذاتي في منمذِج الحزمة الإحصائية للعلوم الاجتماعية.

أولاً: قم بانتقاء عدد النماذج التي ترغب في استعمالها، وإن كنت بصدد بناء عدد كبير من النماذج فسيكون بعضها - على ما يبدو - غير دقيق، وقد لا تريد استعمالها. ونحدد مجال «عدد النماذج التي نريد استعمالها» في 4، مما يعني أننا سنسقط نتائج نموذج واحد. ونرتب بحسب الدقة العامة (أما الاختيار الآخر، فيتجلى في عدد المجالات) لكي نحافظ على أربع نماذج أكثر دقة، فضلاً عن ذلك، نختار الترتيب حسب الدقة في جزء الاختبار بدلاً من مجموعة التدريب حتى يكون بإمكاننا انتقاء مزج النماذج التي تعمم بشكل أفضل على البيانات الخارجية.

ولكن كيف يمكننا تحديد النموذج الأكثر دقة؟ هذا يتوقف على إعدادات التكاليف والعائدات والترجيح. وإن كُلّ ترصد أو سجل «يكلف» النموذج قدراً معيناً من محاولة التصنيف، ويكافئ النموذج «بالعائدات» إن حصل على التصنيف

الصحيح. وفي **الأخير**، يمكن تحديد متغير ترجيح ما، مما يمنح أهمية أكثر لتصنيف بعض الحالات عوضاً عن أخرى، على سبيل المثال، تلك التي «تمثل» أشخاصاً عديدة في ساكنة ما بدلاً من عدد قليل من الأشخاص نسبياً.

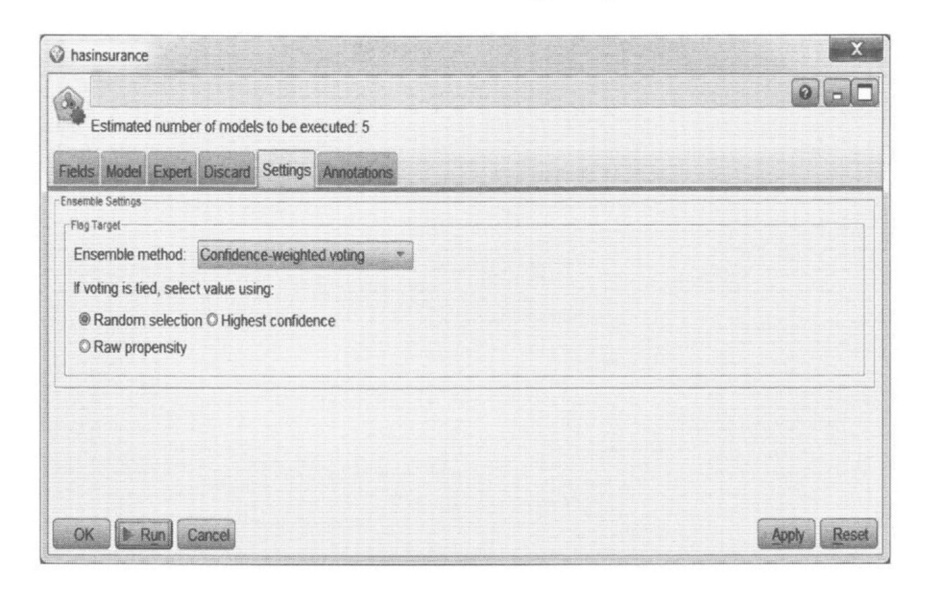


الشكل رقم 15.9: وضع مَعْلمات تحديد الدقة في المصنف الذاتي.

وفي جدولة الطرح (Discard)، يمكن للشخص اختيار طابق من أجل إدراج النموذج. وهذا يعني أن النماذج التي تفشل في الحصول على حدّ أدنى معين من مستوى دقة، اختاره الباحث، ستطرح وإن كانت من أفضل النماذج. وفي الواقع، يمكن للشخص اختيار طوابق مختلفة - مثل نسبة الدقة أو الفائدة الإجمالية، أو المساحة تحت منحنى خاصية التشغيل المتلقي (ROC). ونفضل حذف فقط حالات غير مناسبة بشكل جيد جداً، حيث الدقة أقل من 60٪، والمساحة تحت منحنى خاصية التشغيل المتلقي (ROC) أقل من 0.65.

وأخيراً، قم باختيار الطريقة لمزج «الأصوات» من نماذج مختلفة (جدولة الإعدادات، الشكل رقم 16.9). وتذكر أن النماذج لا تحسب فقط فئة متنبئة لكُلّ حالة، بل أيضاً ثقة في هذا التصنيف. وبعدها، يمكن للشخص تبني النموذج ذي الثقة العالية لكُلّ حالة، أو يمكن للبرنامج القيام بتصويت أغلبية بسيطة، أو يمكنه أن يؤدي تصويتاً مرجح الثقة. واعتبر هذا شبيهاً بإجراء التصويت في تصنيف k-أقرب الجيران. ويمكن للتصويت أن يطرح إشكالية إذا كان هناك عدد زوجي من النماذج، التي ستفصح عن القرار، ولكن المنمذج هنا، يقدم اختياراً بشأن ما يمكن القيام به في حالة تعادلـ اختيار عشوائي أو ثقة أعلى. ونفضل استعمال تصويت الثقة المرجحة.

ويشغل النموذج كلا من النماذج الخمسة قيد البحث وحوسبة دقة كُلّ واحد منها، ثم يطرح الأقل دقة، الذي كان في هذه الحالة الانحدار اللوجيستي. وبعد ذلك، يأخذ القيم المتنبأة بالنسبة إلى مجموعتي الاختبار والتدريب، من كُلّ النماذج المتبقية، ويمزجها من خلال منح كُلّ واحد منها صوتاً مرجحاً لثقة النموذج المقدرة. ويسفر إجراء المزج على مجموعة جديدة من قيم متوقعة، تعد بالأساس التخمينات الأفضل من أصل أفضل أربع نماذج.



الشكل رقم 16.9: اختيار الطرق من أجل مزج متنبئات من مصنفات متعددة.

الجدول رقم 6.9: مقارنة أداء تمازجات مؤمثلة للمصنفات مع المصنفات الفردية.

	الدقة (التدريب)	الدقة (الاختبار)	الحساسية (التدريب)	الحساسية (الاختبار)
الانحدار اللوجيستي	٪72.19	٪73.06	٪73.91	٪72.27
k- أقرب الجيران	٪75.14	٪72.53	٪72.76	٪72.77
آلة متجهة الدعم	٪78.60	٪72.70	٪74.90	٪72.02
الشبكة العصبية	٪72.63	٪73.75	٪76.16	٪71.52
شجرة التقسيم	٪75.35	٪73.19	٪75.61	٪70.95
مصنفات المزيج	٪79.72	٪75.71	٪75.95	٪75.48
المصنفات الممزوجة بالمتغيرات المعالجة سلفاً	٪80.94	٪76.63	٪76.98	٪76.31

وكما يشير الجدول رقم 6.9 إلى ذلك، إن مزج المصنفات يتفوق بالفعل من حيث الأداء، على روتينيات التصنيف الفردي. علاوة على ذلك، يتفوق هذا المزج على جميع المصنفات الفردية على مستوى بيانات التدريب والاختبار، وتقوم بذلك في الوقت الذي تحقق فيه التوازن بين تصنيف الإيجابيات الصادقة والسلبيات الصادقة بشكل صحيح. وتوضح هذه النتائج وجود شيء يكتسب عبر مزج المصنفات، طالما أن الشخص يقوم بذلك بعناية. وبالنتيجة، على الأرجح أن يقوم مزج خمس نماذج سيئة بأداء أسوء من أي نموذج جيد (على رغم من أنه قد يقوم بأداء أفضل من كُلّ النماذج الخمسة من تلقاء نفسها). ومن ناحية أخرى، ليس التحسن مثيراً في هذه الحالة، على الرغم من أنه ملحوظ. وثمة سببان وراء هذا.

أولاً: إنّ كلاً من النماذج الفردية تعمل - سلفاً - عملاً جيداً (وخذ بعين الاعتبار أن التخمين العشوائي قد يؤدي إلى 50٪ من نسبة الخطأ في هذه الحالة).

ثانياً: إننا نستخدم متغيرات متنبئ كبير - إلى حدّ ما - في هذا النموذج (المنطقة، والعمر، ووضع سوق الشغل، ودخل الأسرة، والجنوسة، والعرق، والمواطنة،

والحضور المدرسي، والحالة الاجتماعية)؛ فمتنبئاتنا لا تقدم - ببساطة - للمصنف مادة خاماً من أجل تنبؤ أفضل. وهذه نقطة مهمة تسجل؛ فالتنقيب في البيانات طريقة ذكية للرفع من القوة الحسابية الخام، ولكنه ليس حلاً في حدّ ذاته للبيانات الرديئة، أو المعلومات غير الكافية، أو الخطأ في القياس، كما يمكن للتنقيب في البيانات، تحسين التنبؤ حتى مع البيانات الرديئة، لكن الطريقة الأفضل لتحسين التنبؤ تبقى نفسها ما دام أنها كانت في القرن العشرين - الحصول على بيانات أفضل.

ولكن يمكننا – في الحقيقة – القيام بأفضل من هذا عبر تزويد كُلّ من المصنفات، ببعض المتغيرات التي قد سبق أن حسناها. (انظر الجدول رقم 6.9 مجدداً). ونتوسل هنا بمتغيرات الورقة (التي تمت مناقشتها مسبقاً، تحت استخدام أشجار التقسيم لدراسة التفاعلات، وكذا الدخل، والمجموعات العمرية، التي تم توليدها منها، من شجرة التصنيف المستخدمة من قبل لتوليد تفاعلات معقدة. ومن خلال استخدام هذه المتغيرات، وتلك التي كنا نستعملها في السابق، نكون قادرين على الدفع نحو تحقيق الدقة التنبؤية. وهذا مثير للاهتمام لأننا استخدمنا شجرة التصنيف كواحدة من نماذج التكوين، ونظرياً كان بالإمكان إيجاد ليس فقط المجموعات التي ولّدِنَ من شجرة التقسيم الصغيرة، وإنما مزيداً من مجموعات ذات دقة متناهية. ومع ذلك، إن تغذية الخوارزمية بأكملها من هذه المجموعات، يحسن من قوة تنبؤية عامة.

الفصل العاشر
أشجار التصنيف

إن شجرة التصنيف - كما طورها بريمان (Breiman et al) (1983)، (المعروفة أيضاً باسم شجرة الانحدار (CART)، أو مربع كاي للكشف عن التفاعل التلقائي (CHAID)، أو شجرة القرار، أو شجرة التقسيم) - هي بطرق ما، أداة التنقيب في البيانات النموذجية: بسيطة، وفعالة، وكثيفة الحوسبة، ولا معلمية، وتعتمد على البيانات، بشكل مطلق؛ فهي أولاً وقبل كُلّ شيء، مصنف، تستعمل خصائص المدخل لخلق نموذج يقسم الحالات إلى الفئات ذات قيم مختلفة على مستوى نتيجة ذات أهمية. ولا يهم إن كان متغير النتيجة أو متغيرات المدخل ثنائية، أو فئوية، أو مستمرة؛ فبإمكان شجرة التقسيم معالجتها بأكملها، والتعامل معها بالطريقة نفسها تقريباً. ومع ذلك، تكون أشجار التقسيم أكثر بساطة لدى استعمالها بنتيجة ثنائية، لذا سنركز عليها.

تتوافر أشجار التقسيم (Partition Trees) - باعتبارها مصنفات - على ميزتين إضافيتين مقارنة مع أدوات تقليدية من قبيل الانحدار اللوجيستي؛ فهي:

أولًا: موجهة نحو تنبؤ متغير النتيجة، بدلاً من تقدير المَعْلمات بدقة بالنسبة إلى المتنبئين.

ثانياً: إنها غير مقيدة لتقدير متوسط العلائق؛ بل بدلاً من ذلك، طورت مجموعة جداً معقدة ومحددة من قرارات التصنيف التي تعمل بشكل مختلف بشأن الأجزاء المختلفة من البيانات.

كيف تعمل أشجار التقسيم؟ إنها تشرع في الاشتغال على كُلّ البيانات، وتركز على متغير النتيجة المحدد من لدن الباحث. ويحدد الباحث أيضاً مجموعة من متغيرات المتنبأ المفيدة احتمالاً، في مهمة التصنيف. وتقسم شجرة التقسيم العينة عند كُلّ قيمة لكُلّ متغير مدخل. وفي كُلّ مرة، تحسب مدى كفء هذا التقسيم في فصل حالات بين فئات مختلفة من فئات متغير النتيجة؛ إذ تختار المتغير والقسمة الذين قاما بأداء جيد في مهمة الفصل هذه، مخلفة مجموعتين فرعيتين (أو عقد منحدرة) (Descendant Nodes)، أكثر تجانساً من العينة ككل (العقدة الجذر) (Root Nodes). وتتكرر العملية في كُلّ عقدة من العقد المنحدرة، لتنتج أربع مجموعات، ثم تتكرر في منحدراتها، وهكذا. وتستمر أشجار التصنيف على هذا النحو إلى غاية الحصول على مجموعات متجانسة من الملاحظات المتجانسة تماماً أو بلوغ نقطة توقف معينة.

ويعد هذا الإجراء شبيهاً بنتائج متعددة الفئات. وفي هذه الحالة، تحاول شجرة التقسيم أن تقسم البيانات إلى مجموعات فرعية متجانسة قدر الإمكان، مما يعني في نهاية المطاف - مع الأخذ بعين الاعتبار قدراً كافياً من التقسيمات - أنها ستنتج عقداً يسيطر عليها صنف أو آخر على نحو واضح. وستعمل التقسيمات الأولية في اتّجاه تحقيق هذه الغاية، ولكن ليس من المرجح أن تنتج عقدة متجانسة بشكل مثالي. والأمر نفسه ينطبق - بطبيعة الحال - على نتائج ثنائية أيضاً.

أما بالنسبة إلى النتائج المستمرة، فلا يمكن للإجراء تقسيم الحالات إلى فئات متجانسة، بل ينتج - عوضاً عن ذلك - مجموعات فرعية، حيث قيم متغير النتيجة متماثلة قدر الإمكان (مما ينتج تباينات كبيرة في المعدلات عبر مجموعات فرعية). وهكذا - مع الأخذ بعين الاعتبار تقسيمات متعاقبة - يخلق الإجراء مجموعات فرعية من البيانات حيث التباين على مستوى المتغير التابع مقلص بشكل كبير.

وبما أن الأشجار تعتمد كثيراً على البيانات، فإنها تناقض بشكل كبير التقنيات الإحصائية الكلاسيكية التي تولي الأولوية لاختبار الفرضية. وإن طريقة الشجرة، لا تنتج أي شيء مماثل لمعامل انحدار ما؛ فهي لا تخبرك عما إذا كان متغير ما، متنبأ نتيجة «ذا دلالة»، شبكة متنبئات أخرى. ولهذا السبب، استقبل مجتمع العلوم الإنسانية

أشجار التقسيم بفتور في أحسن الأحوال. (انظر مثلاً، رودجير وآخرين (Ruger el al. 2004; Weerts and Ronca 2009). وتستخدم الأشجار على نحو واسع في ميادين من قبيل علم الأوبئة، والنمذجة الإيكولوجية.

وتعد الأشجار قيمة بالنسبة إلى الباحثين لأسباب ثلاث على الأقل؛ فهي:

أولاً: على ما يبدو أفضل في إنتاج تنبؤات دقيقة من الانحدارات، مثلاً. وإذا ما زودت ببيانات كافية ومتغيرات مستقلة كافية، فستنتج نموذجاً أكثر تناسباً.

ثانياً: فيتمثل في عدم وجود أي حدود بشأن عدد المتغيرات المستقلة التي يمكن إدراجها داخل نموذج ما، ولا توجد صلة هنا بمعضلات درجات الحرية.

ثالثاً: وكما فصلنا القول في ذلك سابقاً – يتمثل في كون أشجار التصنيف جيدة للغاية في إيجاد التفاعلات والعلاقات اللا خطية. ويمكن لنماذج الانحدار فقط التعامل مع الأشكال اللا خطية فقط إذا كانت محددة مسبقاً من لدن الباحث، وتميل التفاعلات في الانحدار إلى الانحصار في متغيرين أو ثلاثة على الأكثر. في المقابل، تولد أشجار التقسيم تفاعلات معقدة آلياً، ومن ثم فهي أداة قوية في البحث الاستكشافي.

وتتدفق إحدى عيوب أشجار التقسيم مباشرة من رحم نقاط القوة هذه، إذ عبر مرونتها وعقدتها، تتمكن الأشجار من توليد نموذج تنبؤي أكثر دقة. ولكن الشجر التي نشأت داخل مجموعة بيانات كبيرة، وتستعمل العديد من المتغيرات، ستكون بالضرورة كبيرة ومعقدة، مما يجعل التأويل عملية صعبة نوعاً ما. وما تجنيه الشجرة من قوة التمييز أو التنبؤ تخسره في التقتير.

عندما تشرف أشجار التقسيم على نهايتها، تواصل تقسيم البيانات حتى لا يتبقى منها سوى عقد طرفية (Terminal Nodes) (أو «أوراق») متجانسة جداً، مع وجود حالات أو ترصدات قليلة جداً في كُلّ واحدة منها. ومع ذلك، يمكن للباحث تحديد قاعدة توقف لمنع هذا التطور. مثلاً، يمكن تحديد حجم أدنى للقسمة. وعليه، فالشجرة لا تقسم عقدة ما إذا كان لأي من العقد الناتجة أقل من عدد معين من الحالات. إن قواعد التوقف مهمة، لأن هدفنا – عادة – ليس ببساطة تصنيف البيانات

الخاصة التي حصل أن فحصناها، وإنما تطوير نموذج يتنبأ جيداً على العموم. ولكن، حتى مع وجود قواعد التوقف في مكانها، تواجه شجرة التقسيم خطراً كبيراً من الإفراط في تناسب النموذج - منتجة نموذجاً متأثراً بشكل مبالغ فيه بخصوصيات البيانات التي بنيت معها، والتي لها صلاحية خارجية قليلة. وللاحتراز من الإفراط في التناسبية واختباره، ينبغي أداء الصلاحية المتبادلة، وعندما يتم أداء الكابح العشوائي تستخدم بيانات التدريب لبناء شجرة ما، وتسقط بيانات الاختبار آنذاك من الشجرة. وإذا كانت البيانات منقسمة إلى ثلاثة أجزاء، فإن مجموعة التدريب تستعمل لزرع شجرة ما، وبعدئذٍ معايرتها أو موالفتها بدقة باستخدام مجموعة الصلاحية. وتُزال (تُشذب) الفروع التي تساهم في عملية الإفراط في التناسبية خاصة، داخل مجموعة الصلاحية، مخلفاً نموذجاً يقبل التعميم على الأرجح. **وأخيراً**، تسقط مجموعة الاختبار الشجرة الموالفة، موالفة دقيقة من أجل اختبار مستقل لدقة النموذج. وبدلاً من ذلك، يمكن أداء الصلاحية المتبادلة لطية k-.

مثل في الغامب برو

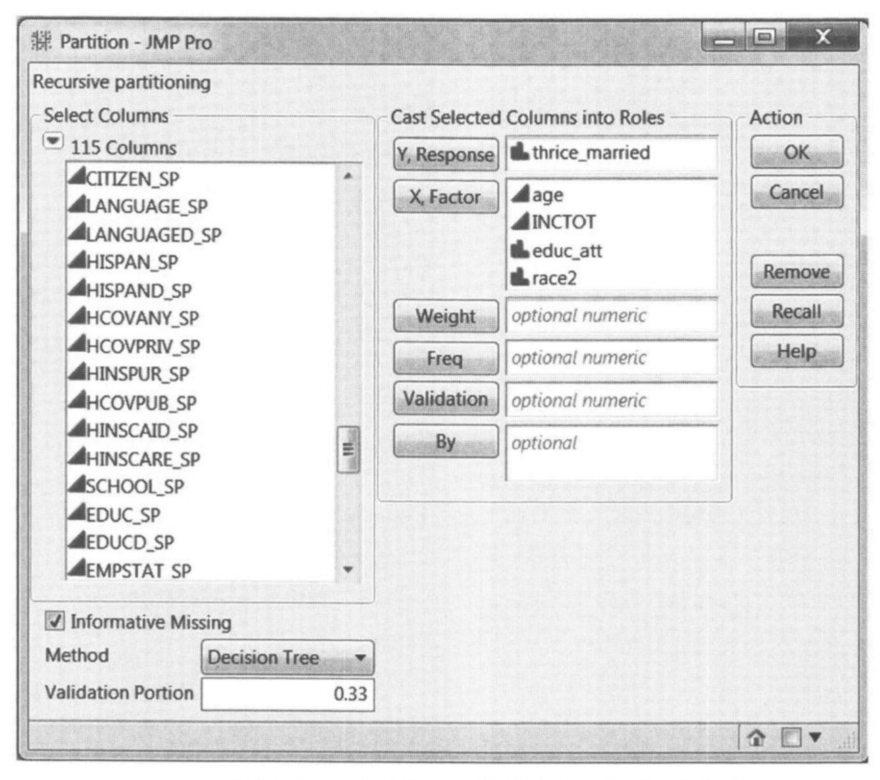
ظلت أشجار التقسيم حاضرة لفترة لا يستهان بها، وحُرّرت روتينيات بالنسبة لـ R، ومنمذِج الحزمة الإحصائية للعلوم الاجتماعية، والحزمة الإحصائية للعلوم الاجتماعية، والستاتا، والماتلاب من بين حزمات برمجية أخرى. وسنعرض الأشجار مستخدمين روتين تقسيم الغامب برو، الذي نستحسنه لسهولة استخدامه، ومرونته، وجودة التصور الذي يقدمه.

سنبين عملية أشجار التصنيف، مستخدمين بيانات من مسح المجتمع الأميركي. وانتقينا عينة فرعية من هذه البيانات التي تحتوي فقط على بالغين تتراوح أعمارهم ما بين 25-65 سنة، وسبق لهم الزواج مرة واحدة على الأقل. وضمن هذه المجموعة، قمنا بمعاينة مفرطة بشكل ملحوظ لتشمل أفراداً سبق لهم الزواج ثلاث مرات أو أكثر. سنستخدم أشجاراً لفصل هؤلاء المتزوجين عن غيرهم في البيانات.

ونفتتح منصة ضبط التقسيم (الشكل رقم 1.10) بانتقاء «حلل التقسيم المنمذج». ونختار كجواب (ويسمى المجال «Y»، جواب) متغيراً وهمياً مشفراً 1، إذا تزوج شخص ما ثلاث مرات أو أكثر، و0 إذا كان الأمر عكس ذلك. وبعدها نختار مجموعة

من المتنبئات (المجال «X»، عامل): السنّ، مجموع الدخل الفردي، التحصيل العلمي، والعرق، والمكانة المهنية، والجنوسة، ومنطقة البلاد، والجنسية/ مكان الولادة (مواطن أميركي بالولادة، أو أميركي مجنّس، أو غير مواطن).

وأخيراً: نختار جزءاً من البيانات لاستبقائها من أجل التثبت من النموذج. ولدينا العديد من الحالات هنا – أكثر من 100.000 – لذا لسنا مطالبين باستخدام الصلاحية المتبادلة لطية k- (على الرغم من عدم وجود أي مانع مبدئياً يمنعنا من القيام بذلك). وبدلًا من ذلك، نحدد حصة الصلاحية في 0.33، مما يبقي على ثلث البيانات من أجل الصلاحية، وإن الشكل رقم 1.10 يظهر ما لدينا قبل الإطلاق.



الشكل رقم 1.10: إطلاق منصة التقسيم في الغامب برو.

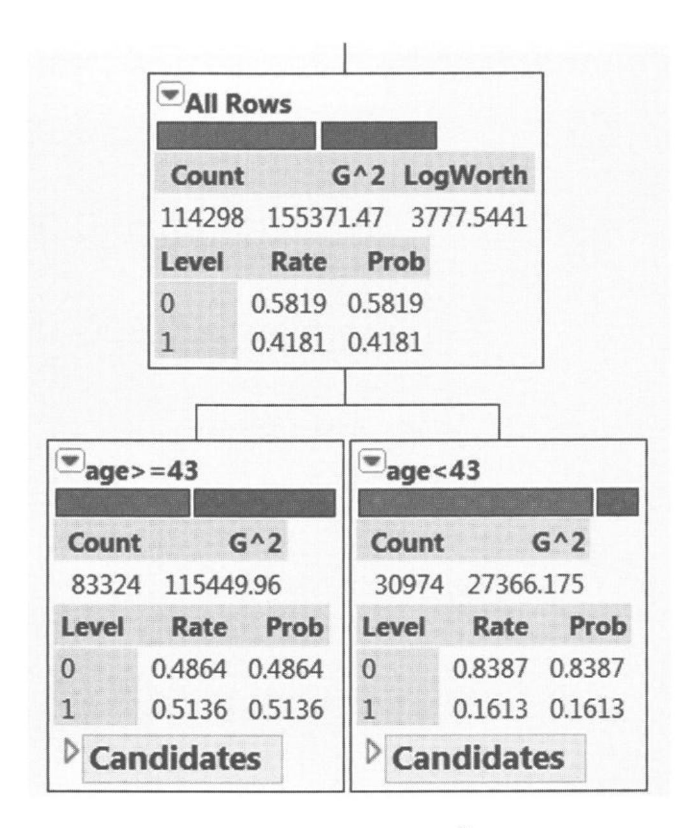
نقر «موافق» (OK)، ونفتح منصة التقسيم، ثم سنقوم بتعديلين لتكييف ما سنراه. ونريد إيجاد النسبة الرقمية للحالات في كُلّ عقدة موجودة في كُلّ فئة من النتيجة (1 أو 0). إذن ننتقي من القائمة (المثلث الأحمر) في الركن الأعلى على

اليمين «أظهر الخيارات» (Display Options)، و«أظهر احتمالية التقسيم». ويبين لنا هذا أن 114.528 من حالاتنا كلها موجودة في عقدة وحيدة (عقدة الجذر) وأن 41.8 بالمائة «متعددي الزواج» (تذكر أننا وسعنا من معاينة هذه المجموعة، حتى لا تعكس معاملات التناسب كميات السكان). وإننا الآن على أتم الاستعداد للقيام بأول تقسيم للبيانات، بنقر الزر الذي يقول «قسّم».

لقد قسمت شجرة التقسيم البيانات إلى عقدتين اثنتين (الشكل رقم 2.10)، بحيث تضم العقدة، إلى اليسار، فقط الأشخاص ممن تصل أعمارهم 43 عاماً أو فما فوق. وأما العقدة الأخرى، فتحتوي على أشخاص أصغر من 43 عاماً. وفي هذا المثال، فقط عقدة واحدة من العقد الناتجة (العقدة الموجودة على اليمين) تعد أكثر تجانساً من منتجها. أما العقدة الأخرى، فهي أقل تجانساً. ولكن على العموم - وعلى مستوى العقدتين معاً - تم رفع التجانس (أو بلغة الأشجار، تم تقليص الأنتروبي (Entropy)) وهذا ما يحاول الإجراء تحقيقه.

وإذا ما أردنا معرفة مكان تقسيم عقدة ما لاحقاً، فإننا ننقر المثلث بجانب المرشحين (Candidates) في أسفل كُلّ عقدة. ويظهر هذا إحصاء القيمة الخوارزمية لكُلّ متغير (أي إن قيمة أو مستوى هذا المتغير الذي يقسم البيانات بشكل أفضل). وسيختار الغامب برو المتغير ذا أكبر إحصاء للقيمة الخوارزمية. وهذه السمة مفيدة، لأنها تسمح لنا بمقارنة متغيرات في كُلّ مرحلة، مشيرة إلى الأجدى منها في تصنيف البيانات.

بعد أن تم إنجاز المزيد من التقسيمات لبعض الشيء، أضحت لدينا صورة أفضل نوعاً ما، مما يميز متعددي الزواج. ونتبع أولاً الفرع الأيسر (الشكل رقم 3.10)، الذي يجد مزيداً من الفرق من بين أولئك الذين تبلغ أعمارهم 43 سنة أو أكثر. ويتم التقسيم الأول في هذه المجموعة، استناداً إلى الميلاد (المتغير cit2)، مع أخذ بعين الاعتبار احتمال زواج الأميركيين الأصليين، بنسبة الضعف أو ثلاث مرات أو أكثر من ذلك، مقارنة بالمهاجرين (بصرف النظر عن وضعية مواطنة المهاجر).

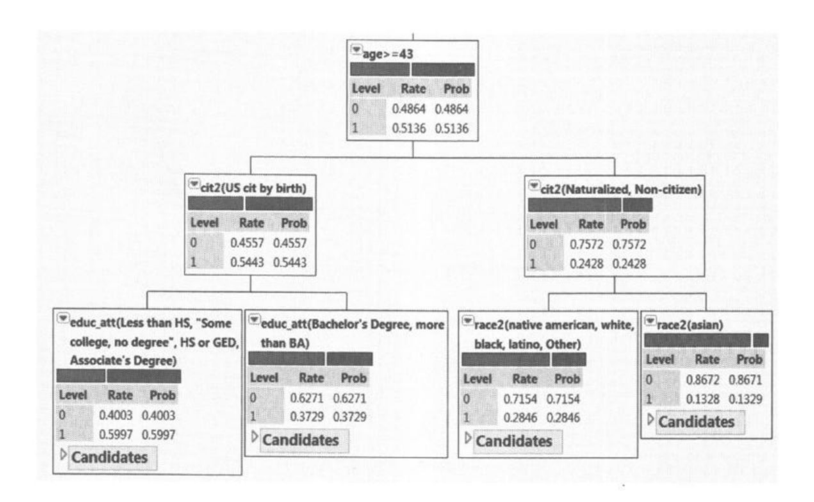


الشكل رقم 2.10: التقسيم الأول للبيانات باستخدام شجرة التقسيم للغامب برو.

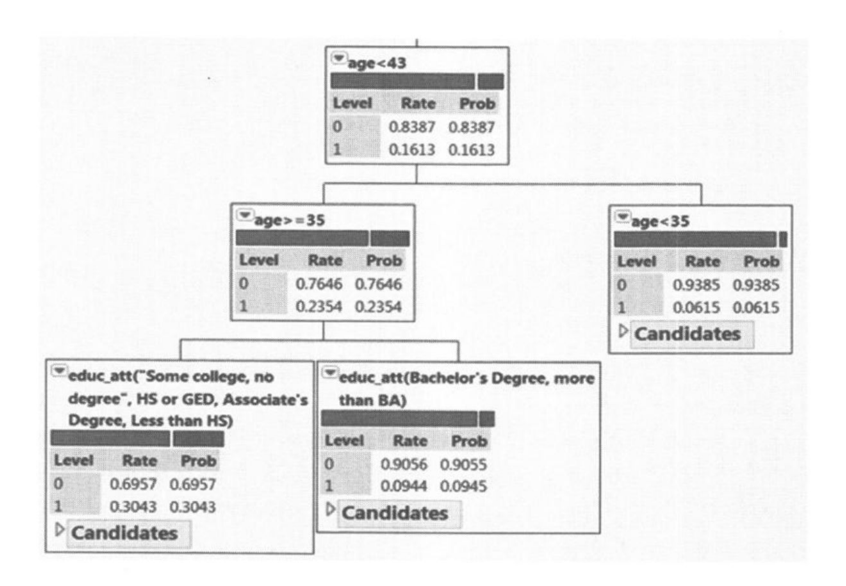
كما ينقسم النموذج من حيث التحصيل العلمي، بين المزدادين الأصليين. أما الأفراد الحاصلين على الباكالوريوس، فلديهم معدل أقل بشكل ملحوظ من الزواج المتعدد، من زملائهم الأقل تأهيلاً. وتم تقسيم من التقسيمات حسب العرق من بين المهاجرين، وأما الآسيويين، فيبدو أنهم من غير المرجح أن يتزوجوا عدة مرات بشكل خاص.

وباتباع الفرع الموجود على الجانب الأيمن (الشكل رقم 4.10)، الذي يحدث تقسيماً بين الأشخاص الأقل سناً من 43 عاماً، نجد قسمة أخرى حسب العمر عند 35 عاماً. ومن غير المرجح جداً أن يقوم أولئك الذين تقل أعمارهم عن 35 عاماً، بالزواج عدة مرات (وتذكر أننا وسعنا بشكل ملحوظ من عينة المتزوجين ثلاث مرات). ومن

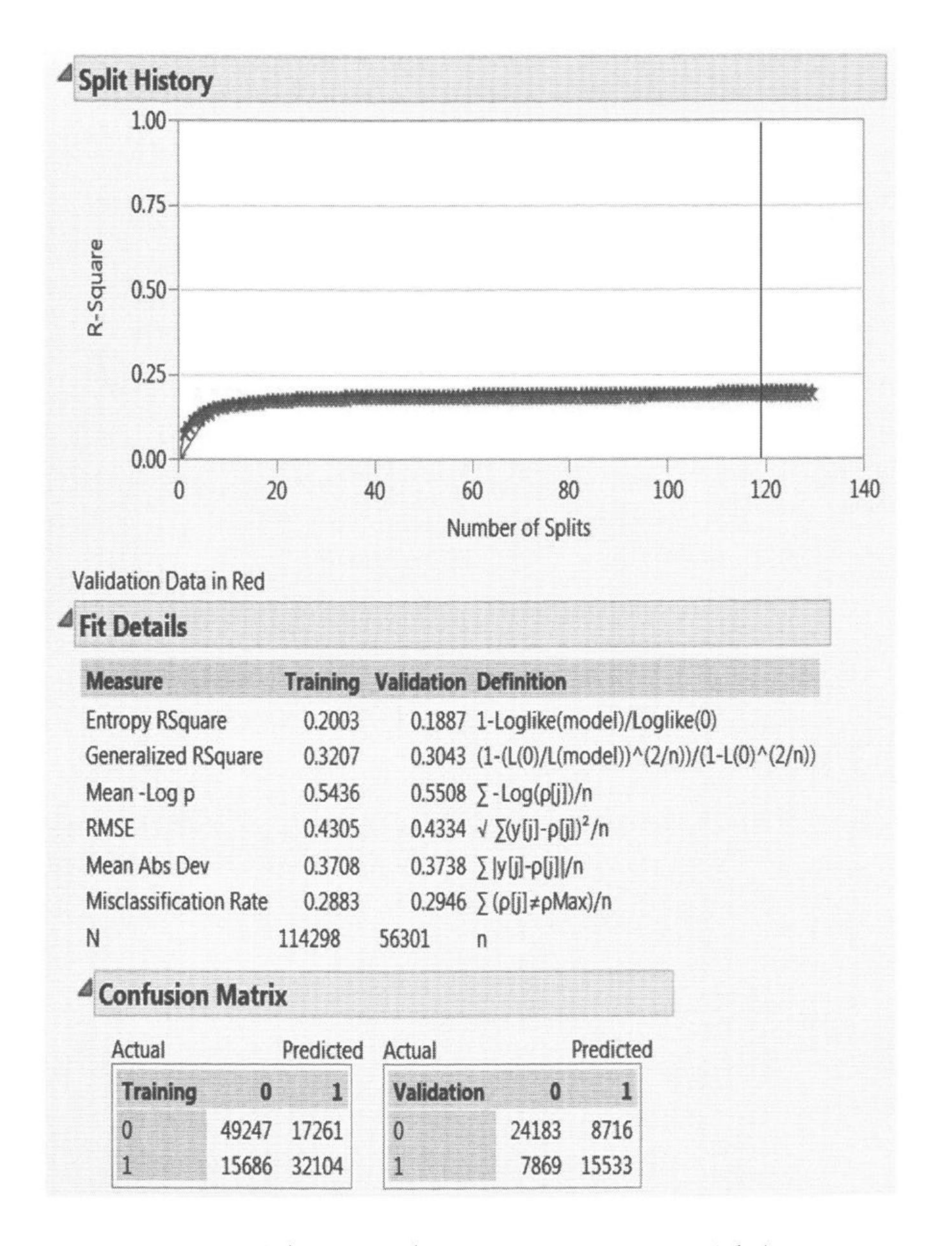
ضمن أولئك الذين تتراوح أعمارهم ما بين 35-42، تقسم الشجرة مجدداً حسب التحصيل العلمي، ومرةً أخرى ليس مرجحاً المشاركة في ثلاث زيجات أو أكثر.



الشكل رقم 3.10: اتباع الفرع الموجود على الجهة اليسرى لشجرة تقسيم ما (الغامب برو).



الشكل رقم 4.10: اتباع الفرع الموجود على الجهة اليمنى لشجرة تقسيم ما (الغامب برو).



Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.2003	0.1887	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3207	0.3043	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5436	0.5508	∑ -Log(ρ[j])/n
RMSE	0.4305	0.4334	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3708	0.3738	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2883	0.2946	∑ (ρ[j]≠ρMax)/n
N	114298	56301	n

Confusion Matrix

Actual / Predicted

Training	0	1
0	49247	17261
1	15686	32104

Actual / Predicted

Validation	0	1
0	24183	8716
1	7869	15533

الشكل رقم 5.10: مخرج شجرة التقسيم في الغامب برو.

إن «الغامب برو» يحسب أيضاً إحصاء تناسبي مشغل، الذي يسميه R^2. وفي الواقع، هذا هو شبه مربع مكفادين (McFadden's pseudo-R^2) الذي يُظهر مدى

تحسن النموذج الحالي مقارنة بالنموذج الصفري أو عقدة الجذر. وفي هذه النقطة، ارتفع القياس إلى 1380. في بيانات التدريب، و0.135 في بيانات الصلاحية.

ولدينا هنا حالات عديدة في كُلّ عقدة، ويمكننا مواصلة القيام بتقسيمات فردية إن كنا نرغب في ذلك. ولكن بدلاً من كُلّ هذا، سوف ننتقل بسرعة إلى الأمام، وننشأ الشجرة برمتها، وذلك بنقر «انطلق».

لقد أنشأ «الغامب برو» شجرة، ويمكن أن نلاحظ (الشكل رقم 5.10) تقسيمه للبيانات إلى 116 مرة، وهذا، دفع شبه مربع «مكفادين» إلى بلوغ 0.189. في مجموعة الصلاحية. ويظهر لنا «الغامب برو» تاريخ التقسيم على مستوى تناسب النموذج. ويظهر هذا منحنيات تحسن متفرقة بالنسبة إلى مجموعة التدريب وإن خط التدريب أعلى قليلاً من خط الصلاحية، لأن القوة التنبؤية هي دائماً أعلى من مجموعة التدريب. ولاحظ أيضاً أنه على الرغم من أن هذين المنحنيين قريبين جداً، فإن الحجم الصِّرف لمجموعة البيانات مقارنة بعدد السمات، ضمن عدم وقوعنا في الإفراط في التناسب بقدر كبير جداً. وتظهر هذه النافذة أيضاً التصحيحات بالنسبة إلى الإفراط في التناسب. ولاحظ الخط الأسود العمودي في 116، الذي هو عدد التقسيمات في الشجرة الأخيرة. وتم إيقاف الشجرة هنا، لأن صلاحية R^2 أعلى مما كان، بعشر تقسيمات إضافية. وبعبارة أخرى، أنجز «الغامب برو» هذه التقسيمات العشر الإضافية، ثم حسِب الـR^2، فاختار الشجرة الأصغر من خلال تشذيب الشجرة الأكبر من أجل تنبؤ مثالي في مجموعة الصلاحية.

للحصول على مزيد من قياسات التناسب، ننقر المثلث الأحمر بجانب «تقسيم المتزوجين ثلاث مرات»، ثم ننقر «أظهر تفاصيل التناسب» (Show Fit Details). وضمن «تفاصيل التناسب»، يمنحنا «الغامب» عدداً من الإحصائيات، بمساعدة الصيغ التي يستخدمها لحسب هذه الإحصائيات. ويوفر صيغتين من R^2، التي يسميها «أنتروبي R^2 وR^2 المعممة» (صيغ مكفادين، وكوكس (Cox)، وسنيل (Snell) على التوالي)، كما يوفر أيضاً متوسط خطأ الجذر التربيعي، ومتوسط الانحراف المطلق، ومعدل سوء التصنيف.

هذا منبر جيد للإشارة إلى مدى بت أشجار التقسيم في «الصنف» التي ينبغي أن

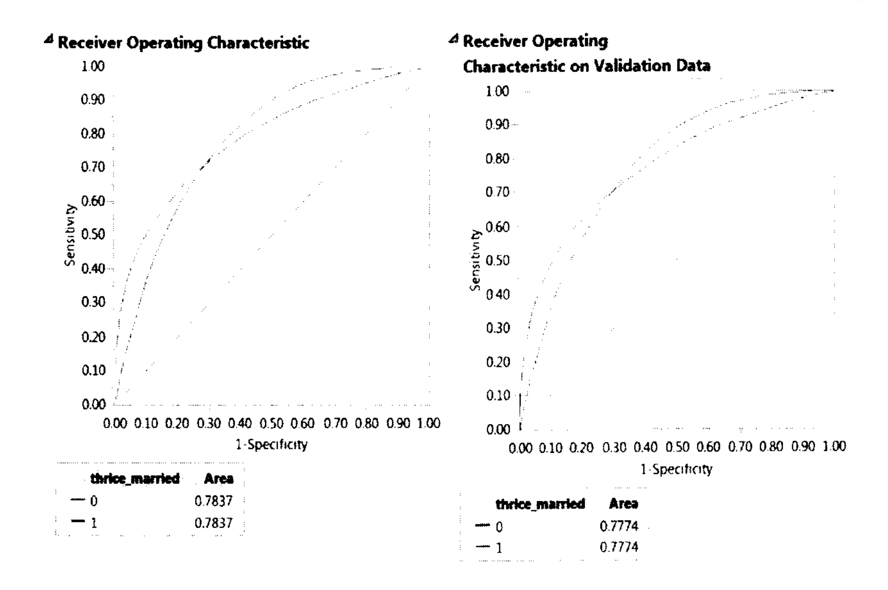
تنتمي إليه حالة ما. لقد سبق أن ناقشنا كيف أن خوارزمية التقسيم ستنشأ شجرة ما إلى حين بلوغ نقطة توقف ما، مما سينتج مجموعة من العقد النهائية (أو «الأوراق»)، التي يحتوي جميعها على حالات من كلا صنفي النتيجة. وأما صنف عضوية المتنبأ به، فتم البت فيه ببساطة عبر تصنيف كُلّ حالة، بصفتها تنتمي إلى صنف النتيجة لغالبية الحالات التي في عقدتها النهائية. وإن نسبة سوء التصنيف هي ببساطة قياس لنسبة الحالات غير الصحيحة المعينة من قبل هذا الإجراء.

يسمح «الغامب برو» أيضاً للباحث بإنتاج منحنيات خاصية التشغيل المتلقي (ROC)، وهي وسائل مفيدة خاصة لتقييم أداء مصنف ما، مثل شجرة التقسيم (الشكل رقم 6.10). فهي تقوم بتخطيط الحساسية أو (معدل الإيجابية الصادقة) بواسطة 1- خصوصية (أو معدل الإيجابية الكاذبة)، مبينين بذلك مدى جودة النموذج عموماً فيما يخص تنبؤ الصنف الذي تندرج ضمنه الحالات[1]. وإن المنطقة أسفل منحنى خاصية التشغيل المتلقي (ROC)، هي مقياس ممتاز للدقة التنبؤية: منطقة من 0.5 تخبرنا عن أن النموذج ليس أفضل في التصنيف من تخمين عشوائي، كما تشير القيم العالية إلى درجة النسبة التي ساعد فيها النموذج في التصنيف.

يقدم لنا كُلّ ذلك فكرة جيدة جداً عن مدى تناسب النموذج للبيانات. ولكن ماذا يمكن للشجرة إخبارنا بشأن متغيرات المتنبئ؟ في هذه المرحلة، نواجه مقايضة بين الدقة التنبؤية، وقابلية التأويل السهلة. لقد أنشأنا شجرة، صنفت بشكل صحيح حوالي 70٪ من الحالات في مجموعة البيانات، وهذا تحسن ملموس بشأن التخمين

(1) إن الطريقة التي نقرأ بها منحنى... هي كالتالي: تصور أن حالات مرتبة من اليسار إلى اليمين ترتيباً يوافق الاحتمال المتنبأ لنتيجة ما، كما تم إنتاجه من قبل نموذج. وكلما تحركنا من اليسار إلى اليمين، فإننا نتحرك بشكل متسق نحو الأسفل في احتمال تنبؤي من أعلى نسبة مئوية إلى أدناها. وفي كُلّ نقطة، تصنف كُلّ الحالات في جهة اليسار باعتبارها إيجابية (على مستوى النتيجة) وكل الحالات في جهة اليمين، تصنف باعتبارها سلبية. ويوضح لنا المنحنى النسب الإيجابية والسلبية التي تصنف بشكل صحيح وغير صحيح من قبل النموذج في كُلّ نسبة مئوية لاحتمال متنبأ. ويمثل الخط القطري في أسفل المركز 50٪. وهذا ما يتم الحصول عليه عبر التخمين العشوائي، أي إذا كان النموذج غير مفيد لمساعدتنا على تصنيف الحالات إلى فئات. كما تمثل المنحنيات الموجودة في أقصى اليسار وفوق هذا الخط، تطورات على مستوى التخمين العشوائي. وهكذا، تظهر منطقة ما تحت المنحنى الأكبر من 0.5، أن النموذج، يمثل عوناً في التنبؤ. وهذا يسمح بالقيام بمقارنة عبر النماذج، ومع مصنفات ثنائية أخرى مثل الانحدار اللوجيستي (المراجع).

العشوائي. لكن للقيام بذلك، ولّدنا شجرة معقدة جداً، وهي الشجرة التي تكونت من 116 تقسيماً متفرقاً. وسيسمح لنا «الغامب برو» بالنظر إلى الشجرة في مجملها في نافذة المُخرج، باستعمال المثلث الأحمر في الركن الأعلى جهة اليمين. اختر «أظهر الخيارات»، «قم بعرض الشجرة». وإن أشجار التقسيم شفافة كلياً، لذا من السهل جداً فهم أي جزء من أجزاء الشجرة. ولكن هذا غير مرضٍ؛ إذ ما نريده في الغالب، هو نوع من أنواع تجميع ما لما يخبرنا به نموذج ما، أي طريقة ما لاستيعاب نتائج النموذج في مجملها، وليس هذا سهلاً باعتبار شساعة الشجرة وتعقيدها، النابعة من داخل بيانات واسعة. وعلى الرغم من كُلّ هذا، من المهم تأكيد أن ما يجعل من أشجار التقسيم أشجاراً تنبؤية للغاية، هي تلك الدقة والتعقيد الذين يجعلانها صعبة الفهم فهماً كاملاً.



الشكل رقم 6.10: منحنيات خاصية التشغيل المتلقي (ROC) باعتباره قياس تناسب نموذج لشجرة التقسيم.

تتمثل إحدى طرق فحص نتائج الأشجار، في فحص محتوى الأوراق نفسها، الذي يمكن القيام به في «الغامب برو»، من خلال فحص تقرير الورقة (تقرير ورقة

المثلث الأحمر). ويعرّف هذا التقرير كُلّ ورقة عبر وضع قائمة بكل التقسيمات التي انخرطت في تشكيلها (والتي تشكل - في الجوهر - متغيرات تفاعل في غاية التعقيد)، وتخبرنا بأن تجزئة الورقة حسب فئات النتيجة. مثلاً، تحتوي ورقة ما، حيث يوجد أفراد متزوجون ثلاثاً، وممثلون في نسبة من أعلى النسب، على أولئك الذين:

- هم مواطنون أميركيون بالولادة.
- هم غير جامعيين.
- يبلغ عمرهم 50 سنة أو أكبر.
- هم من عرق «آخر»، وأميركيون أصليون أو بيض.
- يقيمون في مقاطعة التعداد السكاني لوسط الجنوب الغربي (تكساس، لويزيانا، وغيرهما).

وعلى النقيض من ذلك، إن ورقة ما، حيث متعددو الزيجات غائبون بشكل افتراضي، تحتوي على أشخاص يوصفون:

- بكون أعمارهم تتراوح ما بين 43-51.
- بكونهم مواطنين أميركيين بالولادة.
- بكونهم حاصلين على درجة الباكالوريوس فما فوق.
- بكونهم يعيشون في إنجلترا الجديدة أو ولايات منتصف الأطلسي.
- بكونهم ذكوراً.

والشجرة الصغيرة قد تكون شجرة سهلة التدبير ذات عدد صغير من التقسيمات، فعالة جداً في مساعدتنا «الحصول» على الشجرة. ولكن لدى هذه الشجرة 116 ورقة منفصلة، بحيث تحدث كُلّ واحدة منها العديد من التقسيمات. ومع ذلك، يمكن فحص كُلّ الأوراق، لأنه يساعدنا على إدراك تجاوز منطق التقسيمات.

Validation Data in Red

Column Contributions

Term	Number of Splits	G^2	Portion
age	28	16662.6179	0.5355
educ_att	20	4604.48991	0.1480
REGION	30	4037.34654	0.1298
cit2	5	3668.12284	0.1179
race2	12	1335.50306	0.0429
female	15	438.205303	0.0141
HWSEI	7	318.263861	0.0102
INCTOT	2	51.3524418	0.0017

الشكل رقم 7.10: أهمية المتنبأ في نموذج شجرة التقسيم.

ربما يكون المسلك الأفضل لفهم كيفية بناء الشجرة، هو النظر إلى ما يسميه «الغامب برو» مساهمات العمود (Column Contribution) (مساهمات عمود المثلث الأحمر). ويولد هذا مخطط تقارن متغيرات المدخل حسب مقدار مساهمتها في جعل الأوراق أكثر تجانساً من عقدة الجذر (الشكل رقم 7.10)[2]. وفي لغة الشجرة، هذا مقدار مساهمتها في تقليص الأنتروبي، التي تقاس عبر إحصاء G^2. وإن المتغيرات التي استخدمت بشكل متكرر من قبل الشجرة لتقسيم البيانات، ستحصل إجمالاً على أعلى G^2. ولكن هذا ليس كُلّ ما يعنينا هنا. إن التقسيمات السابقة، التي أفرزت حصصاً أكبر من البيانات، ستكون أهم بالنسبة إلى G^2 من التقسيمات اللاحقة. ولهذا، نلاحظ في الشكل، أن متغير المنطقة استعمل للقيام بمزيد من التقسيمات، أكثر من متغير التحصيل العلمي (20 مقابل 30)، ولكن للتحصيل العلمي قيمة G^2 أعلى. وهذا راجع إلى كون العديد من التقسيمات السابقة الأكثر تبعية، قد استُعملت بالتوسل بالتحصيل العلمي، كما «تفسر أكثر» ما يفرق متعددي الزيجات عن غيرهم من الأفراد الذين لم يتزوجوا بالمرة.

(2) إن المتغيرات التي تظهر في الشكل تشير إلى العمر، والتحصيل العلمي (Educatt)، والتعداد (المنطقة)، والمواطنة/ لأصل (cit2)، والعرق/ الإثنية (عرق 2)، والجنوسة (إناث)، والمكانة المهنية (HWSET)، وإجمالي الدخل الفردي (INCTOT).

إذن بم تخبرنا الشجرة بشأن ما يفرق متعددي الزيجات عن باقي الأشخاص المتزوجين؟ **أولاً**، وهذا غير مفاجئ خاصة، إن السنّ يشكل لحدّ الآن، المساهمة الأكثر أهمية، بحيث يستعمل للقيام بـ 28 قسمة منفصلة، والعديد منها يحدث في مرحلة مبكرة في الشجرة. طبعاً، نحن نعلم أن هذا يتعلق ببساطة بالعرض - الأشخاص الذين عمروا لمدة أطول، كانوا «عرضة لخطر» الزواج لمدة أطول، ومن ثم وجود احتمال أكبر كي يتزوجوا مرات متعددة. إننا بطبيعة الحال، نفترض أن معظم التقسيمات التي تشمل العمر، تفرز مزيداً من متعددي الزيجات في الفريق الأكبر سنّاً. ويمكن التأكد من هذا من خلال الانتقال عبر الشجرة بكاملها، وفحص كُلّ هذه التقسيمات. كما نرى أيضاً قيام التحصيل العلمي بمساهمة مهمة، بحيث تخبرنا نظرة ما إلى تفاصيل الشجرة، عن أن أولئك الذين لهم تحصيل علمي أعلى، هم أقل احتمالاً بكثير فيما يخص زواجهم مرات متعددة؛ وهو أمر مفهوم باعتبار ميل ذوي التعليم العالي إلى الزواج في فترة متأخرة، ويواجهون خطر الطلاق بنسبة أقل. ونرى بعد ذلك، استخدام تلك المنطقة من البلاد في الكثير من التقسيمات. إن الزواج المتعدد أكثر شيوعاً في مناطق الجنوب والجنوب الغربي منه في أماكن مثل إنجلترا الجديدة والساحل الشرقي (على الرغم من الصورة النمطية عن أن كاليفورنيا هي عاصمة الطلاق). **وأخيراً**، إن متغيرنا بخصوص المواطنة والمولد، منقور مرات عديدة. ولكن في أغلب هذه التقسيمات، نجد من غير المرجح وجود ذلك بين المواليد الأجانب خاصة من هو متعدد الزيجات.

ولكن للأسف، إن اتّجاه العلاقة بين متغير مهم والنتيجة، لا يعبر عنه بشكل ميسَّر، على الرغم من إمكانية إنتاج خلاصة عن المتغيرات المهمة في بناء الشجرة. ولن تبعد الأشجار أي شيء واضح جداً كمعامل انحدار للتعبير عن القوة ومنحى علاقة معينة. وليس هذا ببساطة ما تجيده الأشجار تحديداً؛ إن كنت مهتماً بالعلاقات المتوسطة، فإننا نقترح عودتك إلى النماذج المجربة والصحيحة لوحدة الاحتمالية والخوارزمية. وإن أفضل ما يمكنك القيام به حقاً في شجرة ما، هو ما نقوم به أعلاه: ولاحظ المتغيرات المهمة، ثم افحص الشجرة وقدم تقريراً عن ما حدث في أغلب التقسيمات التي تشركها.

خلاصة

إن أشجار التقسيم أدوات قوية للتصنيف والتنبؤ. وقد تم تذييلها بشفافية نتائجها وسهولة فهم خوارزميتها الأساسية؛ فهي كثيفة من حيث الحوسبة، ولكن ليست معقدة خاصة. وفي الحقيقة، إنها تعمل بأداء الكثير من الحسابات البسيطة نسبياً؛ إذ توسم بسهولة التوظيف، والاستخدام مع إجراءات الصلاحية المتبادلة. علاوة على ذلك، تستطيع أن تخبرنا عن المتغيرات الأكثر أهمية في تنبؤات التوليد. وإن نقطة ضعفها هو أنها لا تخبرنا بدقة عن مدى أهمية متغير ما.

وتستخدم الأشجار بشكل واسع، وهي شعبية، ووفّرت عدداً من المتغيرات الأكثر تعقيداً. وأسفله، نفحص اثنتين من هذين «الأشجار - العليا»: الأشجار المعززة، والغابات العشوائية.

الأشجار المعززة والغابات العشوائية

لنقل إنك تظن أن شجرة التقسيم، طريقة مثيرة للاهتمام من خلال طرق النظر إلى تصنيف البيانات. ولكن هذا عمل بسيط جداً. أليس هناك من طريقة تمكننا من أخذ قوى طريقة التقسيم، ولكن قم بتكثيفها كي نستغل بشكل حقيقي، قدرة الحاسوب على استخراج أعداد هائلة من الحسابات؟ إذا أحسست بهذه الطريقة، فإن الإحصائيين قد طوروا جواباً عن دعواتك، وفي الواقع، طوروا العديد منها، غير أننا سنركز هنا على اثنتين منها: الأشجار المعززة، والغابات العشوائية. ويمكن أن نتصور كلاهما باعتبارهما أشجار تقسيم على الستيرويدز (Steroids)، معززة من حيث التعقيد وكثافة الحساب. وسنفسر ماهيتهما، ثم نصف كيف يمكن تشغيلهما في «الغامب برو».

الأشجار المعززة

تستخدم الشجرة المعززة عدداً من أشجار أصغر، للتعلم من أخطاء تصنيف سابق، وبناء نموذج أكثر دقة - وذلك ما نأمله. **أولاً**، تنشأ شجرة واحدة بعدد صغير محدد سلفاً من التقسيمات. ثم تحسب احتمالاً متنبئاً به، وبقايا لكُلّ حالة في مجموع البيانات. ويعاد ترجيح الحالات حسب هذه البقايا، بحيث تتلقى الحالات المصنفة

تصنيفاً سيئاً، ترجيحاً أكثر من تلك المصنفة تصنيفاً صحيحاً (وهي عملية تدعى التعزيز (Boosting)).

Boosted Tree

Gradient-Boosted Trees Specification

Number of Layers:	100
Splits Per Tree:	3
Learning Rate:	0.1
Overfit Penalty:	0.0001
Minimum Size Split:	17

☑ Early Stopping

☑ Multiple Fits over splits and learning rate:

Max Splits Per Tree	12
Max Learning Rate	0.5

OK Cancel

الشكل رقم 8.10: منصة الشجرة المعززة في «الغامب برو».

وتنشأ بعدها، شجرة صغيرة أخرى باستخدام هذه الحالات التي أعيد ترجيحها، ويعاد الإجراء مرات معينة عديدة إلى أن يتم توليد نموذج نهائي. والأشجار المعززة هي إذن نماذج تكرارية قادرة - نظرياً - على التعلم من الأخطاء وعلى أن تصبح تدريجياً أكثر دقة مع الوقت.

ولتشغيل شجرة معززة في «الغامب»، افتح منصة إطلاق التقسيم (حلل تقسيم النمذجة). وبالقرب من الركن الأسفل على اليمين من هذه النافذة، انقر «طريقة» وانتقي «الشجرة المعززة». وبعد ذلك، قم بتوطين ما تبقى من النافذة، كما قد تقوم

بالشيء نفسه مع شجرة التقسيم، ثم انقر «موافق» (OK). وهذا من شأنه فتح منصة إطلاق الشجرة المعززة (الشكل رقم 8.10)، التي ستمكن المستخدم من تكييف عملية التعزيز.

أولاً: نقوم باختيار عدد الطبقات (Layers). وهذا هو عدد الأشجار التي ستكون مناسبة بشكل إضافي للبيانات. وإن لدى «الغامب» نسبة تصل إلى 50. ويمنح مزيد من البرنامج فرصاً أكثر للتعلم والتحسن، ومن ثم سينتج نموذجاً أكثر دقة. وفي المقابل سيزيد من العمل الذي يستوجب على الحاسوب القيام به، وكذا الوقت الذي سيأخذه تشغيل البرنامج. أما بالنسبة إلى مجموعات البيانات الضخمة مثل الذي نستخدمها، فيمكن أن تنتج عدداً طويلاً من التشغيل، وقد ينهار البرنامج بسهولة إذا لم يكن للحاسوب ذاكرة وصول عشوائية (RAM) كافية.

ثانياً: نختار عدد تقسيمات كُلّ شجرة على حدة. ويؤدي مزيد من التقسيمات في أي شجرة، إلى أن تصير أكثر دقة، بما أنها ستولّد عقداً نهائية أصغر وأكثر دقة. (تذكّر أن الشجرة التي تمت مناقشتها سابقاً، استعملت 116 تقسيماً، ولم تفرط في التناسب). مرة أخرى عموماً، إن مزيداً من التقسيمات هو شيء أفضل، ولكن، كما سبق أن ذكرنا، إن مزيداً من الانقسامات يزيد أضعافاً مضاعفة من عدد الحسابات التي يحتاج الحاسوب إلى إنجازها، كما يمكن أيضاً أن يزيد من مقدار الوقت المطلوب.

وبعد ذلك، نحدد معدل التعلم (Learning Rate) الذي يتراوح ما بين 0 و1، بحيث تضمر القيم العليا حاجة البرنامج إلى مزيد من الثقة في استنتاجاته الأولية، بينما المعدلات المنخفضة يرسخ مزيداً من الحذر، وهكذا، إن المعدل العالي للتعلم يسرّع من الحسابات المعنية، ولكن على حساب الإفراط في التناسب، في حين تبطئ المعدلات المنخفضة التقارب، وإن كانت تنتج دقة أكثر.

ثمة مَعْلمان اثنان أيضاً، يعملان كوقاية ضد الإفراط في التناسب في منصة الشجرة المعززة «للغامب برو». وتضمن عقوبة الإفراط في التناسب، عدم وجود الحالات ذات احتمالات منبأة مساوية للصفر. وستنتج القيم العليا إفراطاً أقل في التناسب. ويمكن للباحثين أيضاً تحديد الحجم الأدنى من الانقسام، مما سيمنع البرنامج من تقسيم أي عقدة تكون أدنى من عدد محدد من الحالات التي توجد فيها.

ففي مجموعات بيانات ضخمة، من قبيل تلك التي نحن بصدد استعمالها، (ومن غير المرجح استعمالها)، ولكن يمكن أن تكون مهمة جداً في مجموعات بيانات صغيرة.

أما الخياران الأخيران - وهما إما مشغلان أو موقوفا التشغيل - فيسميان «التوقيف المبكر» (Early Stopping) و«متعدد التناسب على التقسيمات ومعدل التعلم» (Multiple Fits Over Splits and Learning Rate). وإن التوقيف المبكر - هذا إذا ما تم تفعيله - يعطي الإشارة إلى البرنامج بتوقيف عملية التعزيز الإضافية في حال فشل مزيد من التعزيز لتحسين التناسب على مستوى بيانات الصلاحية. وأما «متعدد التناسب على التقسيمات ومعدل التعلم»، فيعطي الإشارة للبرنامج من أجل بناء شجرة معززة منفصلة لكُلّ مزج ممكن من التقسيمات ومعدلات التعلم المحددة من قبل الباحث. (ويتم تعيين الحدود الأكثر انخفاضاً من هذه الكميات في خانات التقسيمات ومعدلات التعلم التي سبق وصفها، في حين يتم تعيين الحدود العليا تحت خانة التأشير في حقلي «الحد الأقصى من التقسيمات لكُلّ شجرة» و«الحد الأقصى من معدل التعلم»). وهذا يسمح لبرنامج الشجرة المعززة من تجريب التمازجات المختلفة لهذه المَعْلمات من أجل العثور على مزيج يعظم التناسب. وإن عملية تشغيله تزيد من فرص العثور على «النموذج الأفضل» (Best Model)، ولكنها تزيد من وقت التشغيل بشكل ملحوظ.

ودعماً لتحليلنا، نختار إنشاء شجرة من 100 طبقة (ضعف القيمة الافتراضية). ونقوم باستخدام التوقيف المبكر، ولكن حددنا أيضاً القيم الدنيا والقصوى لكُلّ من التقسيمات بحسب كُلّ شجرة، ومعدل التعلم، وسمحنا «للغامب برو» باختيار قيم هذه المَعْلمات التي عملت بشكل أفضل في تصنيف الحالات بشكل صحيح في مجموعة الصلاحية. وتراوحت تقسيمات كُلّ شجرة من 3 إلى 12، بينما تراوح معدل التعلم من 0.1 إلى 0.5. وعلى عكس شجرة التقسيم التي أعطت النتائج على الفور، استغرق برنامج الشجرة المعززة بهذه المواصفات حوالي ثمانية دقائق للانتهاء، وذلك غالباً بسبب أننا طلبنا من البرنامج إنشاء عدة أشجار معززة بشكل منفصل.

يبين الشكل رقم 9.10، نتائج الأشجار الثمانية عشر كلها التي أنشأناها في مختلف إعدادات التقسيمات ومعدل التعلم. وقد تم إنتاج الطبقات الـ 100 المحددة كلها للأشجار المعززة ما عدا الشجرتين الأخيرتين؛ بحيث تم اشتغال التوقيف المبكر بالنسبة إليهما، لأن إضافة مزيد من الطبقات كان سيؤدي إلى تناسب سيئ.

وانتهى الروتين باختيار نموذج شجرة معززة ذات عدد منخفض نسبياً من التقسيمات لكُلّ شجرة (5)، ومعدل عال نسبياً من التعلم (0.4). وبالنظر إلى توقف الشجرة المعززة عن إضافة الطبقات في الحدّ الأقصى المحدد لدينا (حتى بالنسبة إلى شجرتنا المنتقاة)، كانت هناك احتمالية تحسين التناسب أكثر قليلاً لو قمنا بتحديد مزيد من الطبقات.

خلاصات تحديد صلاحية النموذج

كان التناسب أدناه الأفضل من بين نماذج التناسب

عدد التقسيمات	عدد الطبقات	معدل التعلم	جذر مربع الأنتروبيا	معدل سوء التصنيف	متوسط خوارزمية p	متوسط خطأ جذر متوسط المربعات	متوسط غياب الخطأ
3	100	0.1	0.1876	0.2946	0.5513	0.4339	0.3853
4	100	0.1	0.1896	0.2960	0.5507	0.4340	0.3833
5	100	0.1	0.1941	0.2908	0.5474	0.4321	0.3794
6	100	0.1	0.1941	0.2919	0.5473	0.4322	0.3783
8	100	0.1	020 0.2	0.2886	0.5420	0.4297	0.3756
10	100	0.1	0.2029	0.2873	0.5419	0.4296	0.3743
3	100	0.2	0.1956	0.2931	0.5466	0.4321	0.3755
4	100	0.2	0.2010	0.2878	0.5423	0.4298	0.3740
5	100	0.2	0.2044	0.2860	0.5404	0.4290	0.3718
6	100	0.2	0.2031	0.2856	0.5410	0.4293	0.3704
8	100	0.2	0.2019	0.2878	0.5426	0.4299	0.3717

0.3703	0.4295	0.5420	0.2875	0.2033	0.2	100	10
0.3696	0.4292	0.5416	0.2877	0.2017	0.4	100	3
0.3686	0.4284	0.5391	0.2849	0.2065	0.4	100	4
0.3677	0.4282	0.5389	0.2844	0.2069	0.4	100	5
0.3679	0.4287	0.5404	0.2853	0.2052	0.4	100	6
0.3676	0.4286	0.5399	0.2865	0.2047	0.4	89	8
0.3671	0.4286	0.5405	0.2860	0.2050	0.4	77	10

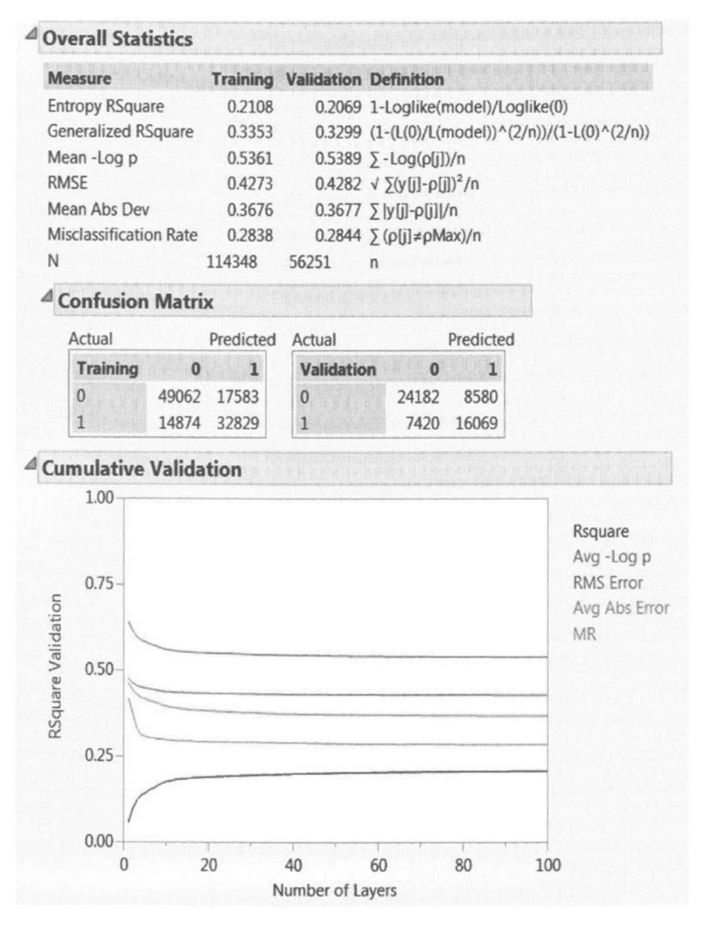
الجدول رقم 9.10: إحصائيات التناسب بالنسبة إلى أشجار معززة متعددة في «الغامب برو».

إن المخرج الناتج عن شجرة معززة (الشكل رقم 10.10)، شبيه بشكل كبير، بمخرج شجرة تقسيم «منتظمة». وإن الفرق الأساسي، هو أن مخطط الصلاحية التراكمية، تبين مقاييس متعددة للتناسب وترسمها ليس في مقابل العدد التراكمي للتقسيمات، وإنما مقابل العدد التراكمي للطبقات أو أشجار التناسب. ويمكننا ملاحظة تحسن أولي سريع في التنبؤ، المنجز من قبل أشجار أولى، وتليها فترة طويلة من تقدم أكثر بطئاً وثابتاً. وبعد الأشجار المعززة، يمكن لمساهمات العمود، ومنحنيات خاصية التشغيل المتلقي، ومنحنيات الرفع أن تولّد أيضاً. وأما تقارير الورقة، فغير متاحة.

يبين لنا مساهمات العمود (الشكل رقم 11.10) مدى أهمية كُلّ مُدخل بالنسبة إلى عملية التصنيف. وبما أن هذه الشجرة مختلفة عن شجرة التقسيم المعيارية – بما أنه تم إنشاء عدد كبير من طبقات الشجرة، كُلّ بحسب بقايا سابقاتها – هناك احتمال أن تكون المساهمة النسبية للمدخلات مختلفة عما رأيناه سابقاً. وبالفعل، هذا هو الأمر الواقع. إننا نرى أن مدخل العرق هو الآن أكثر أهمية من مدخل المنطقة، على الرغم من أن المدخلين العاليين في شجرة التقسيم (العمر، والتحصيل العلمي)

يبقيان المدخلين الأعليين هنا. واستغلت المكانة المهنية - التي بالكاد تلعب دوراً في السابق على الاطلاق - في الغالب من أجل خلق التقسيمات. وتراجعت مساهمات المواطنة، ومكان الولادة المتعلقة بمُدخلات أخرى.

ويمكّن أيضاً «الغامب برو» الباحثين من معاينة الطبقات الفردية. وبحكم صغر حجم كُلّ طبقة، فسيكون من الممكن معاينتها برمتها بسهولة. ومن الممكن أن تظهر الأشجار بدرجات مختلفة من التفاصيل. ولمعرفة أكبر قدراً من المعلومات كما هو مبين في الشكلين رقم 12.10 و13.10، انقر المثلث الأحمر، أظهر الأشجار، أظهر الأسماء، الفئات.



Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.2108	0.2069	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3353	0.3299	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5361	0.5389	∑ -Log(ρ[j])/n
RMSE	0.4273	0.4282	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3676	0.3677	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2838	0.2844	∑ (ρ[j]≠ρMax)/n
N	114348	56251	n

Confusion Matrix

Actual / Predicted

Training	0	1
0	49062	17583
1	14874	32829

Actual / Predicted

Validation	0	1
0	24182	8580
1	7420	16069

الشكل رقم 10.10: مخرج الشجرة المعززة في «الغامب برو».

وقد اخترنا هنا عشوائياً طبقتين (طبقة 8 وطبقة 63) لأغراض توضيحية. وإن إظهار الطبقات الـ 100 كاملة، لن يكون عملياً. وتقوم هاتان الشجرتان بقرارات تقسيم مختلفة تماماً، بحيث تستخدم الأولى في التحصيل العلمي، والعمر، والمكانة المهنية، في حين تستخدم الثانية في العمر، والدخل، ووضع المواطنة.

Column Contributions

Term	Number of Splits	G^2	Portion
age	92	7214730.13	0.2574
educ_att	42	4586316.76	0.1636
cit2	49	4246375.72	0.1515
race2	42	3919236.14	0.1398
REGION	41	3616481.2	0.1290
female	49	1896117.07	0.0677
HWSEI	120	1428779.73	0.0510
INCTOT	65	1117773.13	0.0399

الشكل رقم 11.10: أهمية المتنبأ الناتج عن نموذج شجرة معززة.

غابات عشوائية

تستخدم غابة عشوائية (Random Forests) ما (أو غابة النظام التمهيدي (Bootstrap Forest))، تقنية يمكن من خلالها توليد عدد لا متناهي من العينات العشوائية، انطلاقاً من مجموعات بيانات متناهية؛ فنظام التمهيد كثيف حوسبياً، نقوم بمعاينة بياناتنا بالاستبدال (Replacement) (وهذا مفتاح)، ومن ثم القيام بتوليد مجموعات البيانات المنفصلة المولدة عشوائياً بقدر ما نحتاج إليه. وبما أن تجميع البيانات الأولية الذي تم عشوائياً انطلاقاً من السكان - ومن خلال إعادة معاينة هذه العينة، فإنه «كما لو أننا» نعيد عينة السكان - مع التحذير (القوي) من أن الحالات غير المدرجة في العينة الأولية، لا أمل لها في الانضمام إلى أي من العينات التي أعيد تشكيلها من النظام التمهيد (بينما لدى تلك التي كانت مدرجة في الأول الاحتمالات نفسها كي تكون مدرجة أو تكون غير ذلك، المعاينات التي تمت إعادة تشكيلها).

هكذا، إن النظام التمهيدي يمكّن المحللين الإحصائيين من احتواء - على الأقل جزئياً - مشكلة وجود عينة واحدة عندما يفترض معيار (المعروف أيضاً بالمتكرر) الإحصاء النظري إعادة معينة متكررة. من أجل هذا، كثيراً ما تستعمل باعتبارها طريقة مبدعة للحصول على أخطاء معيارية أكثر «قوة» (Robust).

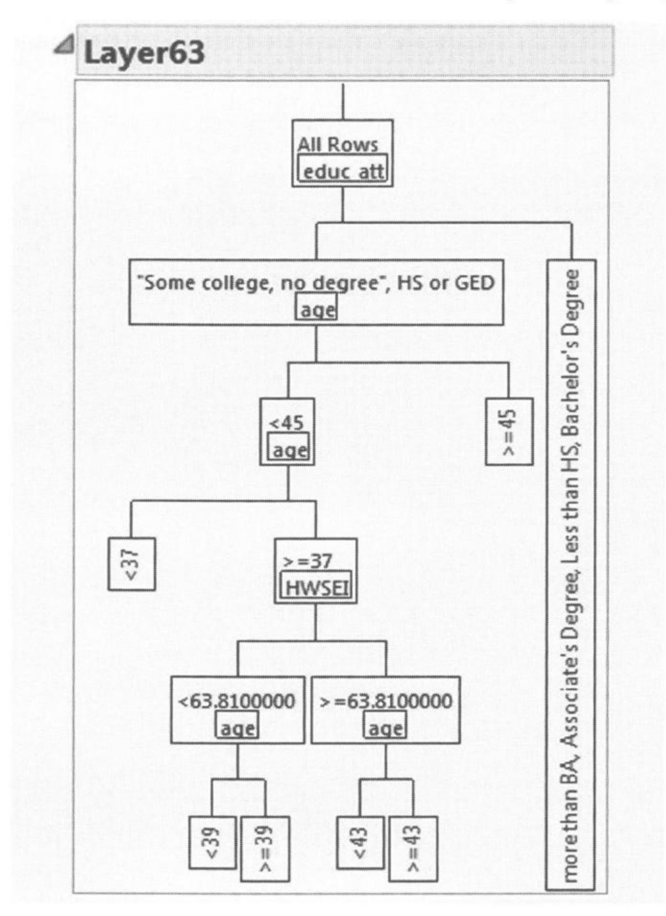
وتستخدم الغابات العشوائية النظام التمهيدي من أجل إنشاء عدد كبير من الأشجار المنفصلة (ومن هنا جاء مصطلح الغابة)، كُلّ واحدة منها يتم على مستوى قسم مختلف من البيانات، المختار عشوائياً (مختار بطبيعة الحال، بالاستبدال). وعلاوة على ذلك، تقوم الغابات بمعاينة قسم من متغيرات المتنبئ، المستخدمة في توليد تقسيمات في شجرة ما. ويضمن هذا أن كُلّ شجرة مولّدة، ستكون مختلفة إلى حدّ كبير. وبعد هذا، سيتم جمع كُلّ الأشجار، واستخراج معدلاتها. والغاية من ذلك شبيهة بغاية الصلاحية المتبادلة: تقليص إمكانية الإفراط في التناسب، والزيادة في احتمالية التعميم.

ولتشغيل غابة نظام التمهيد، قم بفتح نافذة إطلاق التقسيم مثلما تم في السابق، وقم بانتقاء غابة نظام التمهيد في مربع الطريقة. وما عدا ذلك، يبقى نفسه. وعندما تنقر «موافق» (OK)، تفتتح منصة إطلاق غابة نظام التمهيد (الشكل رقم 14.10)، لتسمح بذلك تعديل المَعْلمات.

أولاً: نختار عدد الأشجار التي سوف يتم إنشاؤها لتوليد الغابة. وكما قد تتوقع، إن إنشاء مزيد من الأشجار، يؤدي عموماً إلى نموذج أكثر دقة وقابل للتعميم، ولكن سيزيد في المقابل من وقت التشغيل أيضاً.

وباستطاعتنا الآن تعديل المَعْلمات التي تحدد معدلات معاينة الحالات والمتغيرات (أو إن شئتم الأعمدة والصفوف). ونحدد أولاً «عدد المصطلحات المعينة لكُلّ انقسام». ويشير هذا إلى عدد المتغيرات المستقلة التي تستخدم في كُلّ شجرة. وتقوم غابة نظام التمهيد بمعاينة المتغيرات المستقلة وكذا الحالات بطريقة عشوائية، (أو إن شئت، تقوم بمعاينة كُلّ من الصفوف وأعمدة مصفوفة البيانات).

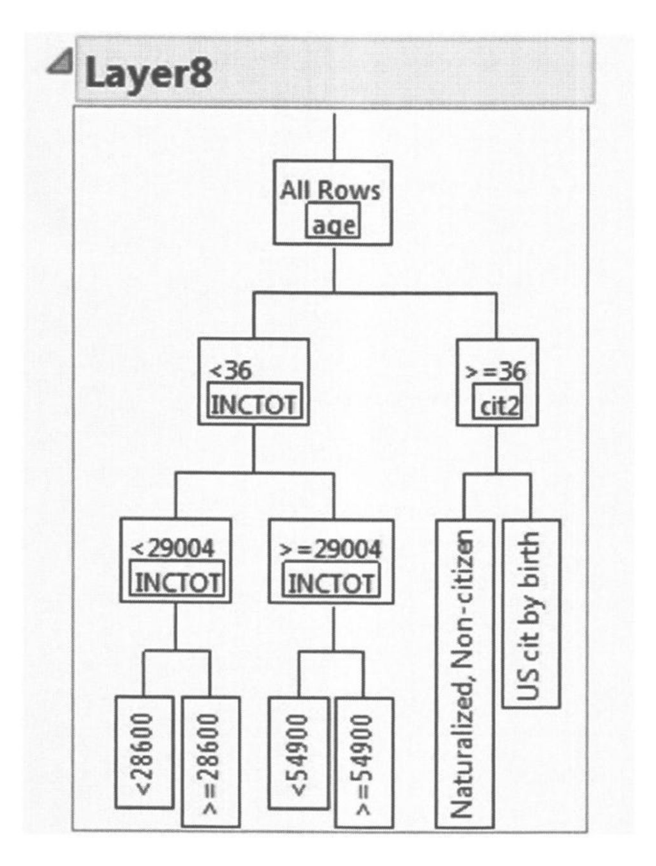
وإن استخدام مزيد من الأعمدة، يسمح لكُلّ شجرة بأن تكون أكثر تعقيداً، ولكن للضرورة، ستكون أكثر تطابقاً، متخلية بذلك عن بعض المزايا من مزايا إنشاء عدد كبير من الأشجار المختلفة وإيجاد متوسط لها. وبعد ذلك، نحدد «معدل عينة نظام التمهيد». ويشير هذا إلى حجم نموذج النظام التمهيدي المراد إنشاؤه من البيانات (المتعلق بنسبة العينة الأصلية).



الشكل رقم 12.10: طبقة واحدة من شجرة معززة.

على سبيل المثال، إن قيمة 100 ستولد نموذج نظام تمهيدي متساوٍ من حيث الحجم مع البيانات الأصلية. هذا الرقم لا يشير الآن إلى نسبة البيانات المستخدمة في نموذج

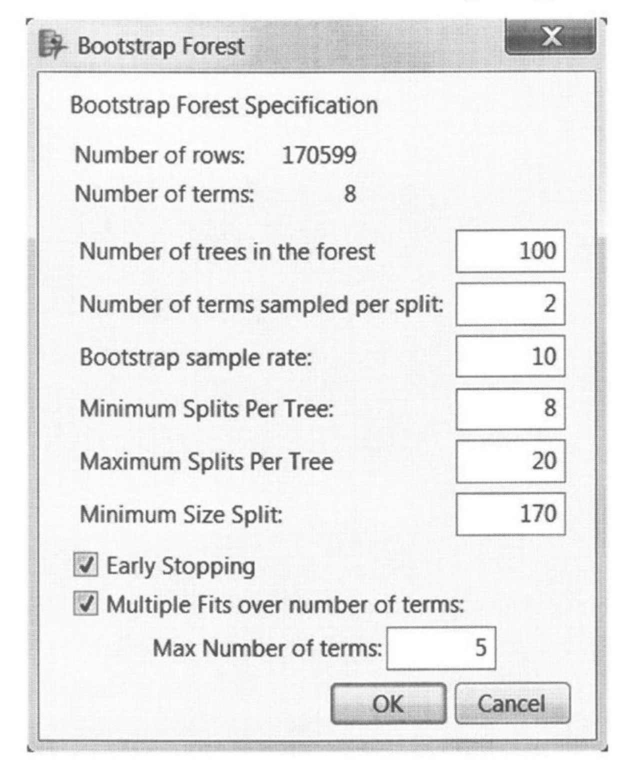
نظام التمهيد، لأننا نقوم بالمعاينة بالاستبدال، ومن ثم يكون من المرجح بالنسبة إلى بعض الحالات أن يتم انتقاؤها أكثر من مرة. ويمكن لهذا العدد أن يكون أكبر من 100. وعموماً، ستؤدي العينات الكبيرة إلى مزيد من الدقة، ولكن أيضاً إلى الزيادة في وقت التشغيل.



الشكل رقم 13.10: طبقة أخرى من شجرة معززة.

وإن «الحد الأدنى من التقسيمات لكُلّ شجرة» (Minimum Splits Per Tree) و«الحد الأدنى من حجم التقسيم» (Minimum Size Split) هما بالضبط ما يظهران: أنهما يعملان من أجل وضع حدود على تعقيد النموذج، ويحذران من البساطة المفرطة، والإفراط في التناسب على التوالي. وكما هو الحال مع الأشجار المعززة،

إن «التوقيف المبكر»، يعطي الإشارة للبرنامج من أجل التوقف عن توليد مزيد من الأشجار، إذا لم تحسن الأشجار الإضافية من صلاحية التناسب. و**أخيراً**، ستقوم «التناسبات المتعددة على مستوى عدد من المتغيرات» - إن تم التحقق منها - بإنشاء غابة منفصلة (Separate Forest) لقيم متنوعة لعدد من المتغيرات، بدء «بعدد المتغيرات التي تمت معاينتها لكُلّ تقسيم»، مروراً بالعدد الذي تم إدخاله في «الحد الأقصى لعدد المتغيرات». ويسمح هذا الخيار بمزيد من النمذجة الشاملة، ولكن يزيد من وقت التشغيل بشكل كبير.



الشكل رقم 14.10: منصة إطلاق غابة نظام التمهيد («البوتسراب») في «الغامب برو».

لقد قمنا بإنشاء غابة من 100 شجرة منفصلة (ضعف القيمة الافتراضية)، واخترنا معدل معاينة نظام التمهيد من 10٪. وتحققنا من «التناسبات المتعددة على مستوى عدد من المتغيرات»، وسمحنا لعدد من المتغيرات لتتراوح تبايناتها ما بين 2 و5، مما

أفضى إلى أن ينتج البرنامج أربع غابات منفصلة مكونة من 100 شجرة لكُلّ واحدة (الشكل رقم 15.10). وبسبب هذا الاختيار الأخير، استغرق البرنامج أربع دقائق للتنفيذ. واستقرت على خمس متغيرات لكُلّ شجرة كعدد مثالي. ومرة أخرى، لاحظ أن الغابة في هذه القيمة المثالية، لا تحتوي إلا على 29 شجرة، مما يعني أن التوقيف المبكر كان شغالاً. ومع ذلك، من الممكن أن تكون معاينة مزيد من المتغيرات قد حسنت من تناسب النموذج.

خلاصات تحديد صلاحية النموذج

كان التناسب أدناه الأفضل من بين نماذج التناسب

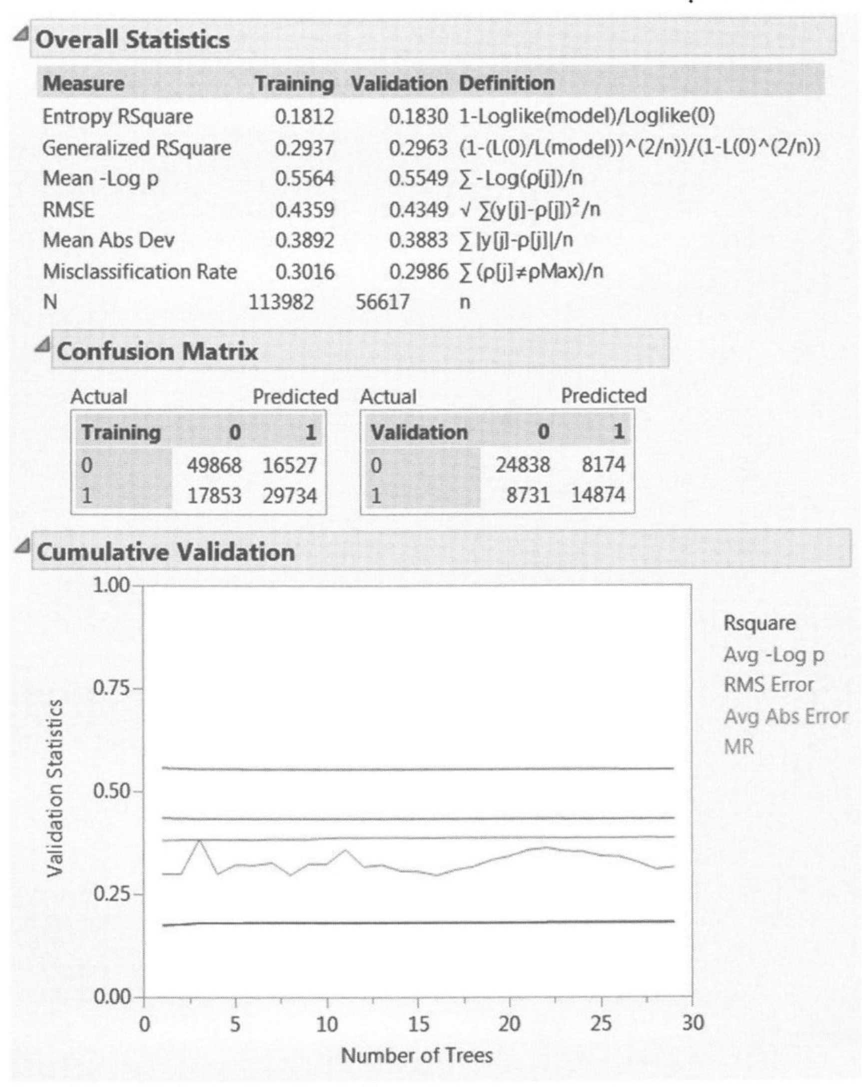
عدد المتغيرات	عدد الأشجار	جذر مربع الأنتروبيا	معدل سوء التصنيف	متوسط خوارزمية p	متوسط خطأ جذر متوسط المربعات	متوسط غياب الخطأ
2	36	0.1579	0.3019	0.5725	0.4420	0.4123
3	100	0.1739	0.3001	0.5612	0.4373	0.3989
4	43	0.1813	0.3014	0.5565	0.4357	0.3905
5	29	0.1830	0.2986	0.5549	0.4349	0.3883

الشكل رقم 15.10: أمثلة نموذج غابة عشوائية في «الغامب برو» من خلال اختيار عدد المتغيرات المعاينة.

لقد تم وصف إحصاءات التناسب ومخرج آخر ذي صلة من الغابة العشوائية في «الغامب»، في الشكل رقم 16.10. وسوف ينتج «الغامب» تلقائياً إحصاءات شاملة للتناسب، ورسم بياني تراكمي للصلاحية مماثل لتلك التي تم إنتاجها بشجرة معززة (باستثناء رسم المحور X لعدد الأشجار في الغابة بدلاً من عدد الطبقات في الشجرة)، ومصفوفة الارتباك. وسوف تنتج أيضاً إحصائيات كُلّ شجرة على حدة.

ومن خلال استخدام المثلث الأحمر في الجانب الأيسر العلوي لنافذة المخرج

الكامل (غير معروض)، يمكننا رؤية «عرض شجرة صغيرة» لكُلّ شجرة فردية في مجموعة البيانات. ويمكننا أيضاً الحصول على معلومات مفيدة مثل مساهمات العمود ومنحنيات خاصية التشغيل المتلقي، ومنحنيات الرفع. كما يمكن توليد الاحتمالات المتنبأة.



Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.1812	0.1830	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2937	0.2963	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5564	0.5549	∑ -Log(ρ[j])/n
RMSE	0.4359	0.4349	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3892	0.3883	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3016	0.2986	∑ (ρ[j]≠ρMax)/n
N	113982	56617	n

Confusion Matrix

Actual / Predicted

Training	0	1
0	49868	16527
1	17853	29734

Actual / Predicted

Validation	0	1
0	24838	8174
1	8731	14874

الشكل رقم 16.10: مخرج من غابة عشوائية في «الغامب برو».

وعند فحص مخطط الصلاحية التراكمية في الشكل رقم 16.10، يكون من

المهم الإشارة إلى أن إحصاءات التناسب التي لا تتحسن بشكل متطابق، بما أن الغابة العشوائية تنشأ مزيداً من الأشجار. وفي المقابل، يوضح مخطط الصلاحية التراكمية لبيان نموذج الشجرة المعززة تحسناً بدائياً سريعاً في التناسب، متبوعاً بتقدم دائم وبطيء. ويرجع هذا الاختلاف إلى الاختلاف فيما تقوم به هاتان الطريقتان في واقع الحال. ولقد صممت الأشجار المعززة للتعلم من الأخطاء السابقة، المؤدية إلى تناسب أكثر قرباً (لكن مع احتمالية الإفراط في التناسب). ومن جهة أخرى، تنشأ الغابات العشوائية أشجاراً فردية بشكل متسلسل، لكن مستقلة عن بعضها بعضاً. وما تقوم به إحدى الأشجار الفردية، هي وظيفة من وظائف المدخلات، والحالات التي تعاينها عشوائياً، وليس وظيفة قامت بها شجرة ما سابقاً. ومن ثم، فالتطور غير مضمون على المدى القريب، على الرغم من أن إحصاءات الصلاحية تعرف تحسناً على المدى الطويل (بينما يتم إنشاء أكبر عدداً من الأشجار، واستخراج متوسطاتها).

إن إحصاءات مساهمة العمود (الشكل رقم 17.10) متشابهة جداً - من حيث العلاقة - مع شجرة التقسيم الأولى. ومرة أخرى، يتم استغلال متغيرات العمر، والتحصيل العلمي، ومنطقة البلد، والأصل في معظم الأحيان لتقسيم البيانات. ومع ذلك، لاحظ أن عدد الانقسامات التي تمت هنا، ضخمة. وهذا راجع إلى أنه في إعدادات الغابة العشوائية، قمنا بوضع الحدّ الأدنى فقط لعدد الانقسامات التي يمكن للنموذج القيام بها بالنسبة إلى كُلّ شجرة على حدة، ولكن لا يتيح أي حدّ أقصى. ولذلك فإن معظم الأشجار في هذه الغابة، هي أشجار جداً دقيقة، وذات انقسامات عدة، لكُلّ واحدة منها.

من الإنصاف التساؤل - في هذه النقطة - عن متغير الشجرة الأفضل من حيث الأداء من أصل المتغيرات الثلاث. أما الأشجار المعززة والغابات العشوائية، فهي متغيرات على مستوى شجرة التقسيم، ولكن كُلّ واحد منها كثيف ودقيق حوسبياً إلى أبعد الحدود. وهل ينجم عن هذا العمل الإضافي نتائج من حيث الدقة التنبؤية الزائدة؟ يقارن (الجدول رقم 1.10) هذه النماذج باستخدام مقاييس مختلفة من التناسب، والجواب الذي يوفره هو نعم، حيث تتفوق كُلّ من الشجرة المعززة والغابة العشوائية في التصنيف خارج العينة. ولكن هل تحسنان النتائج بشكل كبير؟ هذا

يتوقف على قراركم. أما فيما يخصنا، فنميل إلى الجواب الذي يعتمد على مدى أهمية أن تكون دقيقاً بقدر الإمكان. وإذا كانت مشكلة تصنيفكم غير مؤثرة بصورة خاصة، ولكن لديكم كميات ضخمة من البيانات والمتغيرات لتنتج من خلالها، فقد ترغب بالحفاظ على تلك الزيادة الهائلة من وقت التشغيل، وتقوم ببناء شجرة واحدة فقط. ولكن إذا كان لديك مشكلة شديدة التأثير (مثل الفرز بين الخلايا المسرطنة وغير المسرطنة)، فربما قد ترغب في انتظار فترة أطول قليلاً وتكون متأكداً أكثر.

Column Contributions

Term	Number of Splits	G^2	Portion
age	195	410285.743	0.5542
educ_att	59	116084.475	0.1568
cit2	56	79917.7466	0.1080
REGION	130	78181.5372	0.1056
race2	73	43114.0089	0.0582
HWSEI	23	8712.87324	0.0118
INCTOT	9	3134.74896	0.0042
female	6	877.846166	0.0012

الشكل رقم 17.10: أهمية المتنبأ من نموذج غابة عشوائية.

الجدول رقم 1.10: مقارنة أداء شجرة تقسيم، وشجرة معززة، وغابة عشوائية.

	شجرة تقسيم	شجرة معززة	غابة عشوائية
شبه R^2- «مكفادن»	0.189	0.2069	0.181
متوسط الخطأ التربيعي للجذر (الصلاحية)	0.433	0.428	0.435
معدل سوء التصنيف (التدريب)	0.288	0.283	0.301
معدل سوء التصنيف (الصلاحية)	0.292	0.284	0.299
منطقة تحت منحنى خاصية التشغيل المتلقي (الصلاحية)	0.778	0.783	0.752

الحساسية (التدريب)	0.653	0.688	0.625
الحساسية (الصلاحية)	0.649	0.684	0.630
الخصوصية (التدريب)	0.753	0.736	0.751
الخصوصية (الصلاحية)	0.750	0.738	0.752

تسخر الأشجار المعززة والغابات العشوائية المنطق الأساسي للأشجار، ولكن تمزجها مع عمليات التعزيز والنظام التمهيدي في محاولة لتحسين دقة النموذج وتعميمه على العينات المستقلة. أما فيما يخص البيانات والإعدادات الصحيحة، فبإمكانها التفوق على والديها الذي هو شجرة التقسيم، ولكن لا تقم بذلك دائماً في تجربتنا. علاوة على ذلك، تتنازل عن الكثير من امتياز شجرة التقسيم - شفافيتها - بواسطة زيادة التعقيد على نحو ملحوظة. وإن العمل الكبير الذي قد يستهلكه فحص كُلّ من العدد الهائل من الأشجار المنتجة من خلال التعزيز أو في الغابات، هو عمل هائل (على الرغم من أنه ليس مستحيلاً من حيث المبدأ). إنها نماذج أكثر تنبؤاً بشكل حصري من والديها - وليس بالشيء المفيد لفهم ما يجري في عملية التصنيف. ولكن إذا أربكت هذه الطرق محاولات التفسير، فإن ما سنناقشه لاحقاً سيكون أكثر صعوبة. وننتقل بعد ذلك إلى مناقشة طريقة «الصندوق الأسود» بامتياز، المتمثلة في الشبكة العصبية.

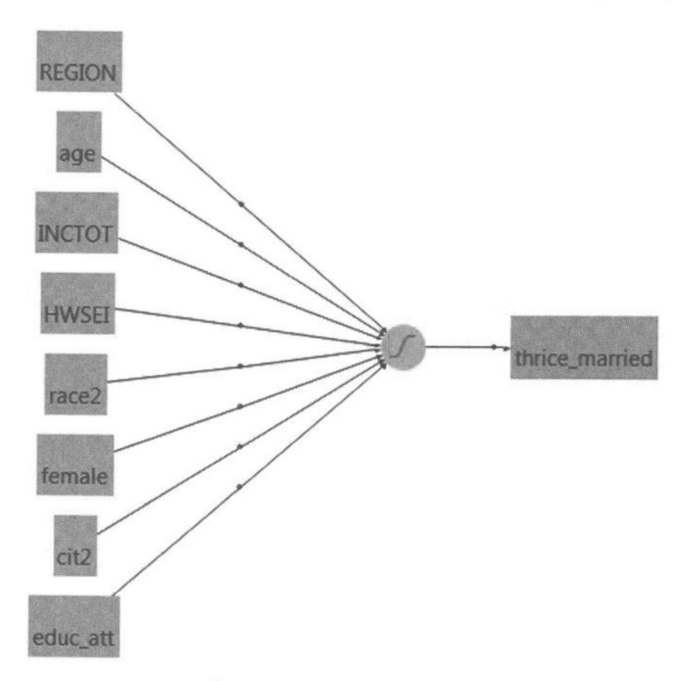
الفصل الحادي عشر
الشبكات العصبية

تعد الشبكات العصبية الاصطناعية (Artificial Neural Networks) (ANN) (التي تعرف اختصاراً، بالشبكات العصبية) أدوات تعليم آلي، تستلهم تقنياتها - كما يقترح ذلك اسمها - من عملية الأعصاب البيولوجية. وللحصول على مفهوم مجرد وعام للغاية، لكيفية اشتغال الشبكات العصبية، لندرس الاشتغال الأساسي لعصب (Neuron) ما. إن لدى الأعصاب تفرعات خلوية (Dendrites) تجمع معلومات مدخلة من أعصاب أخرى. وتدمج هذه المعلومات حتى إذا ما تم بلوغ عتبة ما، «يتقد» (Fires) العصب. وبهذه الطريقة يمدّ العصب قنوات المعلومات لأعصاب أخرى. علاوة على ذلك، تملك شبكات الأعصاب القدرة على التعلم (Learning)، استناداً إلى الأخطاء السابقة.

إن الشبكات العصبية الاصطناعية تعمل على نحو مماثل، ذلك بأنها تجمع معلومات من مجموعة من المُدخَلات (Inputs) (مجموعة بيانات ذات مجموعة معينة من متغيرات مُدخل مستقل). ويخصص لكُلّ متغير مدخل، ترجيحاً عشوائياً، وبعد ذلك تُجمع المعلومات من كُلّ المتغيرات عبر الإجمال (Summation)، وتتحول إلى قيمة نتيجة ما بواسطة دالة لاخطية. ويمكن لمتغيرات المُدخل والمُخرج أن تكون مستمرة، فئوية (Categorical)، أو ثنائية (Binary).

ويظهر مثال من أمثلة الشبكة العصبية كالذي تم وصفه آنفاً في الشكل رقم 1.11. لقد حددنا نموذجاً من ثمانٍ مدخلات، أو متغيرات مستقلة. وتدمج المعلومات من هذه المتغيرات في الدائرة لتُظهر شكلاً مصقولاً يشبه شكل S (Smoothed-S Shape) (ممثّلاً دالة الظل الزائدية (Hyperbolic Tangent Function)، وتستخدم في تنبؤ متغير المدخل.

وسيساعدنا هذا على تقديم بعض مصطلحات الشبكة العصبية. والمستطيلات الثمانية الموجودة على اليسار هي عُقَد المدخلات (Input Nodes) أو طبقة المدخل (Input Layer) للشبكة العصبية التي تقدّر المعلومات، وتلخصها، وتحولها انطلاقاً من المدخلات. و**أخيراً** يشير المستطيل على اليمين إلى طبقة المُخرج (Output Layer) التي تمثل الاحتمال المتنبأ للنتيجة.



الشكل رقم 1.11: شبكة عصبية بسيطة (صورة مأخوذة من «الغامب برو»).

إن هذه شبكة عصبية بسيطة، تعد - بشكل تام تقريباً - انحداراً لوجيستياً. وعموماً، تُحسّن الشبكات العصبية الانحدار اللوجيستي، من خلال إضافة التعقيد عبر عقد خفية متعددة. وتظهر شبكة عصبية أكثر نموذجية في الشكل رقم 2.11،

بحيث تتألف الطبقة الخفية الآن من أربع عقد خفية. وكل متغير في مدخل الطبقة مرتبط - بشكل مستقل - بعقدة خفية، إذ ترتبط بدورها باستجابة المتغير.

ولفهم سبب أهميتها، دعنا ندرس فعل الشبكة العصبية عندما تمرر المعلومات من المدخلات إلى عقدة خفية وحيدة. وتمثل كُلّ عقدة متغيراً وحيداً، له مجموعة محددة لقيم توزيع معين (ثنائي الحدود (Binomial)، وعادي، وهكذا). وبينما يمرر كُلّ متغير معلوماته للعقدة الخفية، يتم تخصيص وزن له، مماثل لمعامل انحدار. وبعد ذلك، تُضاف قيم الترجيح (Weighted Values) بمفردها إلى جانب متغير (اعتراض) وتتحول النتيجة عبر دالة محددة. وينتج هذا قيمة مُخرجَة.

والمهم هنا الذي يجب توضيحه وهو أن الترجيحات المشار إليها آنفاً، اختيرت بشكل عشوائي (Randomly Chosen) من قبل منصة (Platform) الشبكة العصبية. ويتم تعديلها - بعد ذلك - مراراً وتكراراً، كلما تطور النموذج عبر البيانات لتصحيح أخطاء التنبؤ. وتحدث العملية نفسها في كُلّ عُقدة خفية (Each Hidden Node)؛ أي في كُلّ عُقدة، تطَبّق ترجيحات مختلفة، منتقاة عشوائياً، على كُلّ متغير، وتُعدّل بعد ذلك بشكل متكرر. وهكذا فعدد القيم المتنبأة للمتغير التابع، المولّد في كُلّ طبقة خفية، يساوي عدد العُقد الخفية في تلك الطبقة. وتخصص أيضاً لهذه القيم المتنبأة ترجيحاً عشوائياً، كما أن هذه القيم المرجحة معدلة أيضاً بشكل متكرر، وممزوجة لإنتاج احتمال متنبأ للنتيجة.

ومن الممكن - إضافة إلى ذلك - حيازة أكثر من طبقة خفية. ويسمح «الغامب» بناء نماذج من طبقتين خفيتين، ويجمع الخبراء عموماً على أن معظم المشاكل ذات طبقات خفية، تعد كافية. وتستخدم الطبقة الثانية ببساطة، الطبقة الخفية الأولى باعتبارها طبقة مُدخل، كما تنجز عملية الترجيح والنمذجة والتحول نفسها، بالتزامن مع أداء الطبقة الخفية الأولى لمدخلاتها.

وتتطلب عملية التصحيح الترجيحية والمكررة قليلاً من التطوير، وتذكّر أن الترجيحات المخصصة في كُلّ عقدة خفية تشبه معاملات الانحدار. وفي الحقيقة تمت إضافة متغير اعتراض (Intercept) أيضاً، ولهذا، فمن الدقة بمكان، التفكير في

كُلّ عقدة باعتبارها تؤدي بالأساس انحداراً لا خطياً. وتختار الشبكات العصبية معاملات انحدار على نحو أكثر تماثلاً للغاية للانحدار اللوجيستي، الذي يستخدم تقدير احتمال أقصى (Maximum Likelihood Estimation)، من انحدار المربعات الصغرى العادية. ومثلها مثل الاحتمال الأقصى، تبدأ الصيغة العصبية من «تخمين» مختار عشوائياً في أفضل القيم وتعدل نفسها بعد ذلك.

ومع ذلك، وخلافاً للانحدار اللوجيستي، فإنها لا تقوم بهذا مستخدمة كُلّ البيانات، بل تستخدم ترصداً بترصد. وبهذه الطريقة، تكون الشبكة العصبية قادرة على التعلم من «الأخطاء» التنبؤية التي تقع فيها عندما تعالج مجموعة بيانات التدريب قصد صقل مَعْلماتها. عليها الآن أن تعدل بشكل متزامن عدداً كبيراً من المَعْلمات، وذلك بالتحرك في عملية معقدة، من عقدة المخرج إلى كُلّ عقدة من العقد الخفية، ومن ثم، لكُلّ عقدة من عقد المدخل، مُعَدِّلة كُلّ ترجيح على طول الطريق. وخلال هذه العملية «تدرب» الشبكة «نفسها» على التخمين الأفضل في القيمة المتنبأة، القائمة على البيانات التي جاءت من قبل.

تتجلى إحدى ميزات الشبكة العصبية في معالجتها اللا خطية (Nonlinearity)، أفضل بكثير من تقنيات الانحدار العادي، مانحة عقداً كافية بخاصة. وهي قادرة على معالجتها من دون مخرج معين من لدن الباحث. والباحث لا يحتاج إلى القيام بعملية الزيادة في متغيرات تفاعل أو متغيرات محوَّلة (مربعات، تحولات لوغاريثمية وغيرها)؛ إن النموذج نفسه هو الذي سيرسم خريطة لها.

ولكن هذا لا يعني القول إن الشبكات العصبية أصبحت (Automated) بشكل كامل، بل إن هناك معلمات نموذج عديدة تحتاج - من أجل تحسين التنبؤ - إلى ترجيح من لدن الباحث عبر تشغيلات متعددة لشبكة عصبية ما، كما سنرى بعد لحظة. وإن عملية ضبط شبكة عصبية تخضع بقدر كبير، لعملية «التجربة والخطأ» (Trial-and-Error).

وتتجلى ميزة أخرى لهذه التقنية، في قدرة تنبؤية معززة. إنَّ الشبكات العصبية عموماً متفوقة على نماذج الانحدار (أو - نظرياً - حتى على أشجار التصنيف) من

حيث توليد التنبؤات الدقيقة. ومثلها مثل أشجار التصنيف، فهي أيضاً تتعامل مع النتائج المستمرة والفئوية بشكل جيد تماماً.

ولكن الشبكات العصبية لا تخلو من بعض العيوب مثلها في ذلك مثل جميع التقنيات.

أولًا: وقبل كُلّ شيء، تعد الشبكات العصبية رديئة السمعة من حيث إنتاجها للمدخل المبهم بشكل تام تقريباً (فهي غالباً ما يشار إليها باعتبارها طريقة «الصندوق الأسود»). وبخلاف الانحدارات، لا تعمل الشبكات العصبية على تيسير الحديث عن العلاقة بين المدخلات والمخرجات. ومن الممكن التوسل «بالغامب» بغية البحث في الترجيحات أو المعاملات التي تكوّن النموذج، غير أنها – مع ذلك – لا تسلم بتأويل سهل. وهكذا، نواجه المقايضة نفسها بين الدقة التنبؤية وقابلية التأويل كما حدث مع أشجار تقسيم كبيرة. ومع ذلك، فإنه مقارنة مع الشبكة العصبية، تعد شجرة تقسيم كبيرة، نموذجاً سهل القراءة. فمن الممكن – بالمحصلة – قراءة أي فرع لشجرة ما وفهمها. ولكن العقد العصبية تمنحنا الترجيحات المولدة بشكل متكرر بالنسبة إلى عدد كبير من معلمات تفاعلية. وعلى سبيل المثال، ليس من السهل استيعاب معنى ترجيح مساهمة عقدة لمستوى أول خفي في عقدة طبقة ثانية خفية معينة.

ثانياً: إن الشبكات العصبية غير متناسقة إلى حدّ ما، ما دامت تقوم على عملية تعلم تكرارية، تقوم بدورها على تخمينات عشوائية أولية. إنَّ تشغيل برنامج شبكة عصبية في «الغامب» مرتين – على البيانات نفسها – باستخدام المتغيرات نفسها، ذات إعدادات المعلم نفسها، والصلاحية، انطلاقاً من الحالات نفسها، سينتِج نموذجين مختلفين لهما إحصاءات تناسبية، تتنوع بشكل كبير. إنَّ عدم استقرار المنصة العصبية (Neural Platform) يتقلص عندما نستعمل مجموعات بيانات ضخمة، ونماذج أقل تعقيداً، ولكنه يبقى غير مهم.

ثالثاً: إنَّ الشبكات العصبية – ما دامت تتخصص في التنبؤ – لها ميل قوي لتعقيد البيانات. ولكن، يمكن التصدي إليها عبر استعمال الصلاحية المتبادلة. وإذا كانت الإحصائيات التناسبية في مجموعة اختبارك أسوأ بشكل ملحوظ من مجموعة تدريبك، فإن ذلك يعني أنك شكلت نموذجاً معقداً ومحدداً جداً، وعليك إعادة الاتّصال بتعقيد النماذج (وعادة ما يتم تحقيق ذلك من خلال تحديد عقد خفية أقل).

وتوجد الروتينات (Routines) بالنسبة إلى الشبكات العصبية في نمذجة الحزمة الإحصائية للعلوم الاجتماعية (SPSS Modeler)، وR (الشبكة العصبية للحزمة)، وSAS (العملية العصبية) و«الماتلاب» وMATLAB (مختبر المصفوفة). وسنشغل مثالاً في «الغامب»، لمرونته العالية، ولتوفيره أدوات تصور لبيانات ممتازة. وفي مثالنا، نستخدم مرة أخرى بيانات من مسح المجتمع الأميركي. وقد تم تغيير هذه البيانات لتضم فقط البالغين ممن بلغوا سنّ العمل، الذين تم توظيفهم خلال المقابلة، ونقوم بعملية معاينة 5٪ من الحالات (لتسريع عملية البرامج). وسنتنبأ بالدخل الشخصي مستخدمين مجموعة من المتغيرات المشاركة (Covariates).

ولبداية تشكيل النموذج، افتح المنصة الأولية لانتقاء المتغير للشبكات العصبية (تحليل النماذج العصبية (Analyze Modeling Neural)). وفي هذه النافذة، يمكن تحديد المتغيرات المستقلة والتابعة في النموذج العصبي. ونشكل Y ليكون اللوغاريثم الطبيعي للدخل الإجمالي الشخصي. ومتغيرات المتنبأ المختارة هي منطقة من مناطق البلاد، والعمر، والتحصيل التربوي، وافتراض يحدد أولئك المسجلين بصفتهم طلبة في أي مكان، والمواطنة، ومكان الازدياد، والجنوسة، والعرق، وعدد الأسابيع التي اشتغل المبحوث خلالها في العام السابق، وعدد ساعات العمل في الأسبوع في العام السابق. ونشكل أيضاً متغير صلاحية (Validation Variable) إذا ما رغبنا في ذلك. وإن القيام بهذا - في الغالب - أمر جيد إذا ما أردت مقارنة مدى اختلاف حالات الضبط للشبكة العصبية في التدريب نفسه ومجموعات الصلاحية.

نقر OK، فنحصل على منصة إطلاق عصبية (الشكل رقم 3.11)، ونواجه مجموعة كبيرة من الضوابط والمَعْلمات التي نحتاج إلى تشكيلها. وتسمح لنا اللوحة العلوية (Top Panel) لهذه المنصة بتحديد إجراء الصلاحية. كما يمكننا الاختيار بين حصة الكابح (Holdback)، أو السطور المقصية (إذا ما سبق لنا إقصاء السطور، فسيكون ذلك مساوٍ لاستخدام متغير صلاحية كابح)؛ أو الصلاحية المتبادلة لطية-ك (k-Fold). ونختار كبح ثلث البيانات للتحقق من صحتها.

أما اللوحة الموالية، فتسمح لنا باختيار عدد الطبقات الخفية التي نريدها، وعدد العقد بحسب كُلّ طبقة، ونوع التحولات (أو التفعيلات ((Activations))، التي نريدها أن تكون في كُلّ عقدة. ولا توجد هنا إلا طبقتان اثنتان. وإن استعمال المزيد من الطبقات والعقد بالنسبة إلى كُلّ طبقة أضعافاً مضاعفة، يزيد من عقدة النموذج، مما قد ينتج تنبؤاً أكثر دقة في مجموعة التدريب، ولكنه يزيد أيضاً من احتمال الإفراط في التدريب في مجموعة الاختبار.

ويستخدم «الغامب» دالة التان (Tanh) التماسية القطعية (Hyperbolic Tangent Function) بصفتها قيمة فرضية (Default) هنا. وهذه دالة سينية (Sigmoidal) (ذات شكل S)، الشبيهة بالدالة اللوجيستية، ولكنها ممركزة ومقاسة. كما يستخدم التفعيل «الخطي» دالة الربط البسيطة للذاتية (Identity) الخطية التي يستخدمها انحدار المربعات الصغرى (Ordinary Least Squares Regression). وأخيراً، هناك التحول الغاوسي (Gaussian) الذي يستخدم دالة e^{-x2} لتحويل المزج الخطي لـ x's. و«التان»، والتفعيلات الغاوسية كلاهما، يسمحان للنموذج بضبط لا خطيات معقدة في البيانات. وإذا استعملنا فقط دالة التفعيل الخطي، فسننجز – في الحقيقة – انحداراً خطياً معقداً.

تسمح لنا اللوحة الموالية باستخدام التعزيز الإضافي إلى الشبكة العصبية، وهذا يعمل بقدر كبير مثل أشجار التعزيز. ونقوم بمواءمة سلسلة من شبكات عصبية صغيرة، الواحدة تلو الأخرى، بحيث تقوم كُلّ شبكة على مخلفات مقاسة مستخلصة من النموذج السابق. ولا بُدَّ لهذه العملية – نظرياً – من أنْ تعمل على تحسين التنبؤ. ويخبرنا معدل التعليم النموذج بالنسبة التي يجب أنْ تُعدّل بها الترجيحات، استناداً إلى معلومات حديثة محصل عليها من النموذج السابق. وإنَّ معدلات التعليم الأكثر انخفاضاً تخفض من معلومات جديدة، وتدمجها أكثر مع تقديرات أقدم. وتنتج معدلات التعليم الأكثر ارتفاعاً (القريبة من 1) ترجيحاً أكبر لبيانات جديدة.

الشكل رقم 3.11: منصة إطلاق عصبية في برنامج الغامب برو.

توجد هنا مقايضة بين سرعة التوافق، والميل إلى التفريط في التناسب. كما تسمح معدلات التعليم العليا بتوافق أسرع، ولكن من الأرجح أن تتناسب بخاصة مع البيانات الخاصة التي يشتغل عليها المرء.

وبعد ذلك، توجد سلسلة من الخيارات المضبوطة ضبطاً دقيقاً. ويشير «تحويل

المتغيرات المشتركة» إلى تحول آلي، يستطيع «الغامب» إحداثة في متغيرات المدخل لتصحيح الانحراف، ومن ثم العمل على «تطبيع» المتغيرات. ويمكن هذا أن يجعل الشبكات العصبية أكثر دقة، ويوصى به على هذا الأساس. **ثانياً**: يتم توفير خيار «تناسب قوي» (Robust Fit) بالنسبة إلى النتائج المستمرة. وهذا يقلص تأثير الحالات النشاز في البيانات. لقد سبق لنا أن قمنا بتسجيل الدخل الذي يضع حدّاً لهذا المشكل، ولكن لا نفرّط في الخيار تحسباً لأي طارئ. ومع ذلك تبقى «طريقة الجزاء» (Penalty Method) طريقة أخرى للاحتراس من الإفراط في التدريب في البيانات من خلال فرض «معلم الجزاء» على التقديرات. ونحدد هنا الشكل الوظيفي لهذا المعلم (ويمكن الوصول إلى قيمة المعلم نفسها بواسطة الصلاحية)؛ والقيمة الفرضية (Default) هي مربع معلم الجزاء، التي نستخدمها إلا في الحالتين التاليتين:

أ. وجود عدد كبير من المتنبئات.

ب. الاعتقاد في أن بعضها أكثر تأثيراً بكثير من غيرها في النموذج.

وفي هذه الحالة، يُنصح باستخدام إما الشكل المطلق، أو الشكل المتلاشي للترجيح (Weight Decay Form).

تستخدم الشبكات العصبية قيماً أولى مولّدة عشوائياً للبداية في عملية تناسبية البيانات، وتعديلها مع مرور الوقت. كما يعطي ضبط «عدد الدورات» (Number of Tours) للبرنامج، تعليمات لإنتاج عدد من الشبكات العصبية المنفصلة (Separate Neural Nets)، مستخدماً قيماً أولى عشوائية مختلفة للترجيحات. ومن هذه الشبكات، ستتم عملية اختيار النموذج الأنسب لبيانات الصلاحية. وبسبب عدم استقرار النماذج العصبية المشار إليها أعلاه، سيكون هذا خيار جيد اتّخاذه. وعلى الرغم من أن ذلك سيزيد من وقت التشغيل، إلا أنه لا بُدَّ للنماذج المتعددة من اشتغالها دائماً من أجل الحصول على تناسب جيد.

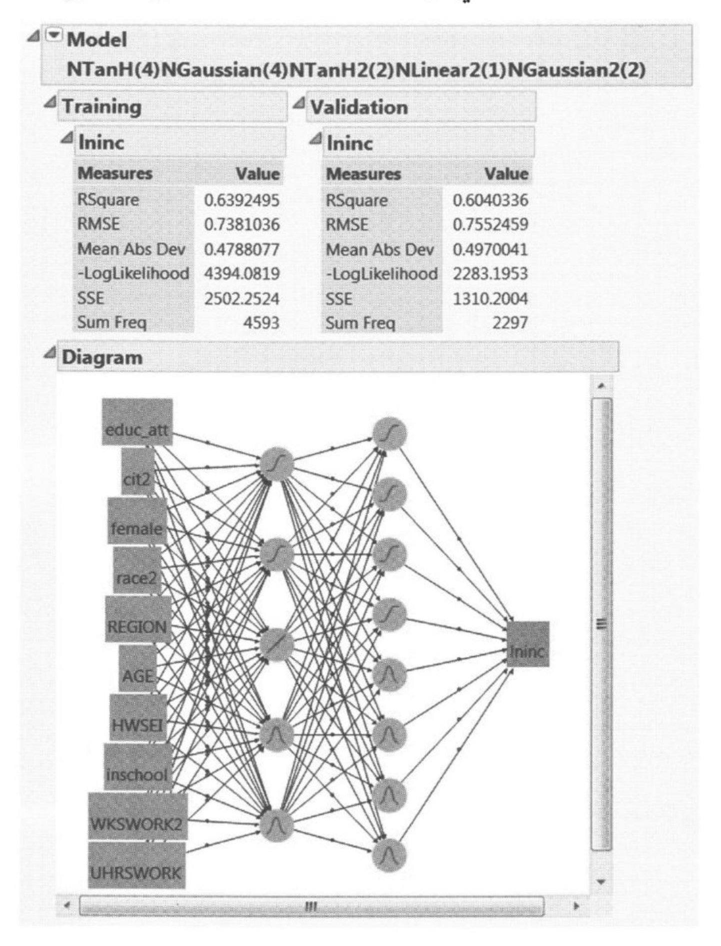
وبعد تشغيل شبكة عصبية، تظل قيمة المخرج المفترضة من «الغامب» ضئيلة جداً (الشكل رقم 4.11)، إذ تتألف، ببساطة، من إحصائيات التناسب. وتعادل R^2 من شبكة عصبية ذات نتائج مستمرة، تحديداً، نسبة من نسب R^2 من انحدار المربعات

الصغرى: إنها نسبة التباين في النتيجة التي يبرزها النموذج. وعلى نحو مماثل، تعد R^2 المنتَجة من أجل الشبكات العصبية ذات نتائج ثنائية أو فئوية مطابقة لنسبة شبه مربع (R^2) المحسوبة من أجل نماذج وحدة احتمالية (Probit) أو لوغاريثمية (Logit) (وفي هذه الحالة شبه مربع مكفادين McFadden's Pseudo-R^2). ويتم تقديم إحصائيات الصلاحية لكُلّ من نموذجي التدريب والاختبار. وعلى المرء مراقبة الفوارق بخاصة في التناسب بين هذين النموذجين من أجل تحديد ما إن كان تناسب النموذج مفرطاً بشكل كبير؛ فإذا كان نموذج ما مفرطاً في التناسب فإن ذلك يعني - عادة - أن نموذجاً إضافياً سيكون أكثر تناسباً مع مجموعة الصلاحية.

ويتفوق «الغامب» في تصور البيانات، كما لا تستثنى الشبكات العصبية من هذا. وإن إحدى خيارات القائمة (في مثلث القائمة بجانب النموذج) وهو الرسم البياني (Diagram) الذي سيقدم تمثيلاً بصرياً للشبكة العصبية الذي شغلتها منذ قليل. ولاحظ وجود ثلاث رموز مختلفة، تظهر في عُقد طبقة خفية. وتشير هذه الرموز إلى دالات التفعيل الثلاثة التي استخدمناها في هذا النموذج من نماذج الشبكة العصبية.

إن عملية نقر المثلث الأحمر في أعلى اليسار نافذة المخرج (الشكل رقم 4.11 بجانب النموذج)، وانتقاء «إظهار تقديرات» (Show Estimates)، سيعرض مكافئ تقديرات المعامل (Coefficient) بالنسبة إلى كُلّ المَعْلمات (Parameters) في النموذج. وإن القيام بذلك، يبين العتمة (Opacity) الشهيرة للشبكة العصبية المشار إليها أعلاه. وعلى الرغم من عملية التعليم التكرارية التي تسعى إلى بنائه، فإنه بإمكان الشبكة العصبية المحصل عليها - مثلها مثل انحدار ما - أن تمثل بصفتها معادلة معقدة وحيدة. وفي الأخير إن ما تم القيام به هو تقدير مجموعة من المَعْلمات. ومع ذلك، فإن ذلك يمثل مجموعة كبيرة من المَعْلمات؛ ذلك بأن الشبكة العصبية أعلاه، مثلاً، قدرت حوالي 200 منها. ويمثل العديد منها كميات من قبيل العلاقة بين العقدة الثالثة الخفية في الطبقة الأولى، والعقدة الخفية الخامسة في الطبقة الثانية. وتعد هذه - طبعاً - نتيجة الخلاصات المتحولة لمدخلات ترجيحية قبلية، وبالتالي لا يمكن تأويلها بشكل منعزل. ويدمج هذا المعلم بمعلمات أخرى، ويخضع لتحول رياضي، ليصير تأثيره في متغير النتيجة أكثر غموضاً. ويصدق هذا على كُلّ المَعْلمات في

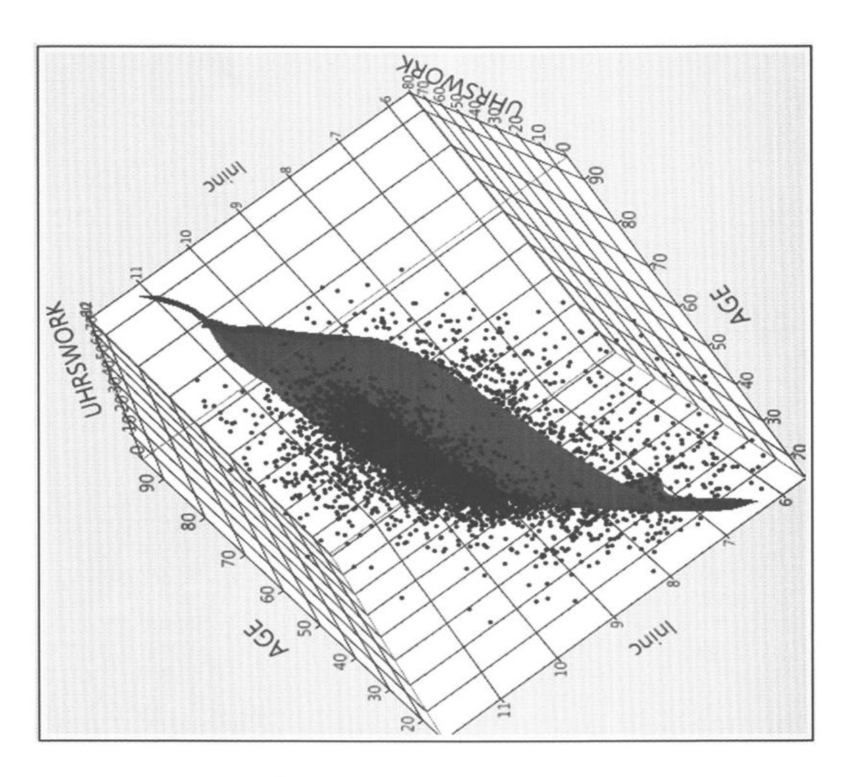
النموذج. ولا معلم من هذه المَعْلمات يملك معنى خارج ظهوره الخاص في الهندسة الكاملة للشبكة العصبية. ولهذا السبب، وعلى الرغم من أن الشبكات العصبية - بمعنى رياضي - شفافة بشكل كامل (Completely Transparent) (لإمكانية التعبير عنها كمعادلة)، فإن العلاقات التي ترسمها مستعصية جداً على التأويل.



الشكل رقم 4.11: مدخل من شبكة عصبية في «الغامب برو» (مع تصور ما للشبكة).

ومع ذلك، فإن «الغامب» ضمّ سمة، تساعد على معالجة العيب. وكما تقوم بذلك مع نماذج الانحدار، فهي تقدم مجموعة من سمات «المحلل» (Profiler)، التي تمكن الباحث من استكشاف العلاقات الهامشية بين المدخلات المتنوعة والمُخرج.

ويمكّننا «محلل السطح» (Surface Profiler) من البحث في تمثيل ذي ثلاثة أبعاد للبيانات (الشكل رقم 5.11). وسيكون محور واحد - الذي هو Z بطبيعة الحال- دائماً متغير النتيجة. أما المحوران الآخران، فيمكننا وضعهما، وهذا يسمح لنا بفحص العلاقة ذات الاتّجاه الثلاثي بين متغيرين اثنين ومتغير الاستجابة بالنسبة إلى إعدادات متنوعة لكُلّ المتغيرات الأخرى. ويمكن تحريك هذا الصندوق تحريكاً ثلاثي الأبعاد، ليتسنى لنا رؤية زوايا متنوعة للعلاقات اللا خطية. وهذه تمثيلات مصقولة للعلاقات المتنبأة من النموذج. وللحصول على فكرة حول كيفية تناسبية البيانات الحقيقية مع هذا، اختر الخيار الحقيقي البارز (Actual Option Under Appearance). وسيرسم هذا نقاط البيانات الحقيقية في الفضاء الثلاثي الأبعاد، إلى جانب العلاقة المتنبأة.



الشكل رقم 5.11: التوصيف الثلاثي الأبعاد للبيانات

باستخدام «الغامب برو» لمحلل السطح.

الجدول رقم 1.11: الدخل المتنبأ، استناداً إلى العرق والجنوسة، المحسوب باستخدام محلل «الغامب برو».

	رجال	نساء	الفرق
أبيض	$43,480	35,950$	$7,530
أسود	$41,360	$29,143	$12,217
لاتيني	$38,560	$32,532	$6,028
آسيوي	$46,630	$34,540	$12,090

ويسمح لنا المحلل (Profiler) البحث في كيفية تأثير كُلّ متغير - عندما يتحرك على طول مداه - في العلاقة بين كُلّ المتغيرات الأخرى والنتيجة. كما يسمح لنا هذا الخيار تحديداً، بالاطّلاع على مدى نجاعة الشبكات العصبية في رسم خريطة اللا خطيات المعقدة في البيانات. كما تمكننا أيضاً من رصد التأثيرات الهامشية بشكل واضح للغاية. ويمكن للباحث وضع كُلّ المتغيرات الأخرى في كميات الفائدة، وبعد ذلك تبديل متغير فردي على طول مداها. يمكننا ذلك مثلاً، من رؤية تأثير العرق، والجنوسة، والدخل الشخصي، وتقييم الفوارق الجنوسية في الدخل استناداً إلى العرق. وبين الأشخاص البالغين سنّ 35 في منطقة الجنوب الأطلسي (الأكبر)، المزاولين لعمل بدوام كامل (40 ساعة في الأسبوع، 50-52 أسبوعاً في العام)، هناك من ولد في الولايات المتحدة، ولهم هيبة مهنية متوسطة (40)، وفي الفئة المتوسطة للتحصيل التربوي بالنسبة إلى السكّان (كلية ما، انعدام الشهادة أو الدرجة العلمية)، قمنا بحساب القيم المتنبأة للدخل استناداً إلى العرق والجنوسة (الجدول رقم 1.11).

تذكر أننا بصدد تثبيت معظم المحددات القوية الحقيقية للدخل (ساعات وأسابيع العمل، والعمر، والوظيفة، والتعليم)، الذي من خلاله يعبر مساوئ سوق العمل عادة عن نفسه. ويقدم لنا هذا - بدمجه مع حقيقة حوزتنا لنسبة R^2 تقدر بـ60. في بيانات الصلاحية - سبباً وجيهاً لأن نكون واثقين من أننا نشهد فوارق حقيقة على مستوى العرق والجنوسة، عوض خطأ المواصفات. وقد سمح لنا «الغامب» من رؤية نمط معقد بشكل واضح لتحديد مشترك للدخل استناداً إلى العرق والجنوسة، من

دون تحديده بشكل واضح في النموذج. وإن تفاعلات من هذا القبيل، تتولّد بشكل آلي بواسطة نماذج الشبكة العصبية.

وتعد الشبكات العصبية لوغاريثمات مألوفة ومرنة للغاية بالنسبة إلى التنبؤ، بحيث يمكن استعمالها في تنبؤ نتائج مستمرة، وثنائية، وذات فئة متعددة؛ وتقوم بهذا بدقة متناهية. كما تستعصي على التأويل بشكل مألوف، على الرغم من أنها تنتج كميات، شبيهة بشكل مباشر بمعاملات الانحدار. ومع ذلك، وبدمجها بسمات من قبيل محلل «الغامب برو»، يمكن استخدامها لفحص علاقات هامشية مهمة، ولو أنه في الوقت الحاضر، لا يمكن إنتاج متوسط التأثيرات الهامشية.

وفي الفصلين المتتاليين، سننتقل إلى فحص سلسلة من الطرق غير المراقَبة من أجل دراسة العلاقات في البيانات.

الفصل الثاني عشر
التجميع

تم استحداث تحليل التجميع (Clustering) لمعالجة حالة مألوفة جداً في البحث. قد تظن أن حالات في بياناتك - مدن، طلبة، أطفال، أو نقابات العمال - لا تمثل عدداً متناثراً عشوائياً بسيطاً من الترصّدات الفردية، ولكنها تصف بشكل أفضل، باعتبارها مجموعات ترصدات. وما نريد القيام به، هو فصل حالاتنا إلى فئات أو تجميعات (Clusters) من الحالات؛ أي القيام بما يشير - بمعنى من المعاني - إلى النوع البسيط والطبيعي جداً من النمذجة الاجتماعية، أي النوع الذي يقوم به كُلّ واحد بشكل ثابت، وعلى أساس مخصص في حياة اجتماعية منتظمة. ولكننا نريد القيام به بدقة، وتطور نظري، ودعم تجريبي، أكثر مما يتم القيام به على نحو طبيعي.

كيف يتسنى لنا - إذن - تشكيل تجميعاتنا؟ وكيف يتسنى لنا تأكيد أن التجميعات التي نشترطها، هي الأفضل - في الحقيقة - أو حتى طريقة لائقة لتصنيف بياناتنا؟ إجمالاً، نحن نسترشد بالنظرية، ونسبة من الترصد: تذكر أنواع طلبة بول ويليس (Paul Willis) (1977)، في كتابه **تعلم العمل** (*Learning to Labour*) أو تصنيف إيسبينغ أندرسون (Esping - Andersen) (1990)، لأنظمة الرعاية الاجتماعية في كتابه **العوالم الثلاثة لرأسمالية الرعاية الاجتماعية** (*The Three Worlds of Welfare Capitalism*)؛ أو إذا كان لدينا ميل أكثر إلى التحليل الكمي، ستقترح - ربما - طريقة من طرق جمع حالاتنا التي تستخدم متغيرين أو ثلاث متغيرات، وبعدها نبحث عن

التثبت من أن الحالات داخل تجميع ما، هي مماثلة - في الحقيقة - من حيث متغير نتيجة ما ذي أهمية (من خلال استخدام أنوفا (ANOVA)، أي تحليل التباين أو الانحدار، لمتغيرات وهمية (Dummy)، على سبيل المثال).

ويمكن اعتبار تحليل التجميع أكثر قوة، وطريقة متطورة من طرق التوجه نحو إنتاج فئات، وتأكيد وجود فئات. ولكن تقوم بذلك من خلال التأثير ليس فقط في بعد أو بعدين من الخصائص، ولكن في أكبر عدد ممكن تحتويه بياناتك، وتراه ذا صلة؛ وكما أن تحليل التجميع، «يؤكد» وجود هذه الفئات عبر استخدام كُلّ المتغيرات المحددة، وليس فقط عبر استخدام هدف متميز أو متغير نتيجة. **وأخيراً** يسمح هذا التحليل بلعب هذه البيانات التجريبية دوراً كبيراً في توليد الفئات، عوض خضوعها لهيمنة النظرية (على الرغم من أن النظرية، تلعب دائماً، دوراً من الأدوار).

التماثل والمسافة

نقوم بتوليد الفئات في تحليل التجميع من خلال تجميع الحالات معاً، التي تعد مماثلة بحسب مجموعة محددة سلفاً من الميزات المناسبة، المشكلة لمتغيرات المُدخل (Input Variables) بالنسبة إلى روتين التجميع. والآن، ما الذي يشكل التماثل (Similarity)؟ رياضياً، يكون ترصدان اثنان أكثر مماثلة، إذا كان لديهما قيماً متماثلة بالنسبة إلى عدد كبير من المتغيرات المُدخلة المحددة، أو بالنسبة إلى جميعها. وهذا أمر بديهي إذا كان لدينا متغير واحد فقط، ولكن التفكير فيه يصبح أكثر صعوبة عندما تكون لدينا مجموعة كبيرة من المتغيرات. ويدخل هذا ضمن مسألة تحديد مفهوم المسافة (Distance) في الفضاء المتعدد الأبعاد (Multidimension Space).

إن لدى الرياضيين طرقاً عديدة لوصف المسافة (لكن لحسن الحظ إن الطرق الأكثر شيوعاً التي تحسب بها المسافة في تحليل التجميع، مألوفة لدينا جميعاً ممن درس الهندسة في الثانوية: المسافة الإقليدية (Euclidean Distance). دعنا نقول إن لديك بعدين ونقطتين في هذين البعدين، ونريد معرفة المسافة بين هذين النقطتين. الجواب السهل عن هذا، هو أن هذه المسافة تقدم بواسطة الخطّ المستقيم (Straight

(Line) الأقصر بين هذين النقطتين؛ ففي الهندسة التي درسناها في التعليم الثانوي، رسمنا النقطتين كليهما على مستوى ديكارتي (Cartesian Plane). وبعد ذلك استخدمنا نظرية فيثاغورس (Pythagorean Theorem) لإيجاد طول الخطّ الأقصر الذي يربطهما، أي بين نقطتين B وA الذين يُحدّد كُلّ واحد منهما بإحداثيين اثنين (x, y) نجد المسافة من خلالـ

$$d_{EUC}(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

ولكن ماذا لو كان لدينا أكثر من بعدين؟ إن الشيء الرائع هو أن هذا الأمر لا يهم - بحيث تعمم هذه الطريقة على ثلاثة أبعاد، وعشرة أبعاد، وn بُعد. وستكون للحالات المماثلة مسافات إقليدية، صغيرة تفصلها، بغض النظر عن عدد الأبعاد المحددة، وإذا ما أردنا معرفة - المسافة - ما بين نقطتين B وA في حيز (Space) محدد بأربعة إحداثيات (x_n, y, z, d)، فسيتم تحقيق ذلك من خلال ما يلي:

$$d_{EUC}(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 + (d_A - d_B)^2}$$

كما يمكن أيضاً استخدام أنواع أخرى من المسافات. وقد تستخدم مسافة مانهاتن (Manhattan) أو («مجمع المدينة» (City-Block))، التي تعد مجموع القيم المطلقة للفوارق بين قيم المدخل. أو نستطيع استخدام مسافة مينكوفسكي (Minkowski) التي تعد تعميماً لكُلّ من مسافتي «إقليدس» ومانهاتن على سلطات عليا؛ أو نستطيع أخذ ارتباطات متغير (Variable Correlations) بعين الاعتبار مستخدمين مسافة ماهالانوبيس (Mahalanobis).

نقاط القوة العامة للتجميع

يمكن أن يكون تحليل التجميع مثمراً بشكل كبير لغايات استكشافية وتوكيدية (Confirmatory)؛ ففي الحالة الاستكشافية، يبدو أن ليس لدينا فكرة ثابتة بعد بشأن

المجموعات الفرعية التي قد تصنف داخلها الترصدات، أو ما إن كان بإمكان هذه المجموعات أن تقسم بشكل مثمر. إن التجميع يسمح لنا بالبحث عن المجموعات الكامنة في البيانات مع الآخذ بعين الاعتبار الخصائص الهامة، كما يمكن أن يخبرنا - بطريقة أو بأخرى - ما إن كانت بياناتنا مجمعة أصلاً، وإذا كان الأمر كذلك، فكيف تختلف هذه التجميعات الأساسية؟ ومن ناحية أخرى، ربما لدينا نظرية ما حول المجموعات الفرعية القائمة. وفي تلك الحالة، يمكن استخدام التجميع لتحديد ما إن كانت نظريتنا مدعمة تجريبياً من قبل بياناتنا وما مدى الدعم. ربما هناك طريقة أفضل(أو ربما سنستكشفها) لعملية تصنيف حالاتنا إلى فئات.

هناك استخدام آخر ممكن للتجمع ويتمثل في استكشاف بنيات التغاير (Covariance) المختلفة في أجزاء مختلفة من البيانات. وإجمالاً عندما نحلل علاقات بين المتغيرات - في انحدار خطي، مثلاً نبحث عن علاقات خطية قائمة في البيانات برمتها؛ أو في الغالب، نولّد متغيرات تفاعل قليلة لبلوغ إمكانية أن تقوم هذه العلاقات على متغيرات أخرى. إن التجميع يسمح لنا بالذهاب إلى أبعد من ذلك؛ إذ عبر التجميع يمكن إيجاد حيز فرعي من البيانات، تكون العلاقات فيه بين متغيرين هو 78. في تجميع A، مثلاً، و24. - في تجميع B. وهذا يعني أن العلاقة بين المتغيرات مختلفة في قطع مختلفة من بياناتنا، وأنه بإمكاننا (استخدام تجميعات لتوليد مجموعات معقدة من متغيرات التفاعل لإدراجها لاحقاً في نماذج انحدار. ويشبه هذا الاستخدام للتجميع - بشكل كبير - نمذجة المزيج، الذي سنناقشه لاحقاً).

نظرية الاعتماد

مع ذلك، من المهم التركيز على أهمية اختيار المدخلات في تحديد الفئات. إن تحليل التجميع ليس وصفة سحرية للكشف عن التجمعات (Groupings) في العالم؛ بل إن ما ينتجه بشكل كامل، هو دلالة من دلالات ما يضعه الباحث فيه. وسواء تشابهت حالتان أم اختلفتا من حيث مساهمتهما الإقليدية، فإن ذلك يتوقف على المتغيرات المحددة، وإذا ما غيرت هذين المتغيرين، فستغير آلياً، المسافات بين الحالات، وفي نهاية المطاف - شكل التجميعات التي تظهر في النهاية. ومن الأساسي الاختيار بعناية الأبعاد التي تضنها مهمة في سؤال بحثك، وتمثيلها جيداً في

مجموعة متغيرات المدخل. وإن تحليل التجميع - في هذه الحالة - شبيه بالتحليل العاملي (Factor Analysis)، وتحليل المكوّن الرئيسي (Principal Component Analysis) الذي يحدد فيه اختيار المدخلات، المكوّنات المحددة أو العوامل الناتجة.

تجدر الإشارة إلى أن لكُلّ متغير في تحليلات التجميع، تأثيراً مماثلاً في تشكيل التجميعات. ومهم أخذ هذا الأمر بعين الاعتبار لسببين:

السبب الأولـ فيتمثل في احتمال أنك لا تظن أن كُلّ متغير يجب يكون مهماً بشكل مماثل لاعتبارات نظرية. وربما تريد أن تكون بعض المتغيرات أكثر ترجيحاًبشكل كبير من غيرها.

السبب الثاني: أحياناً، تستخدم أكثر من متغير واحد لالتقاط بعد معين أو حقيقة اجتماعية؛ إذ يمكن التعبير عن التشكيل العرقي لمدينة ما - مثلاً - فقط عبر مجموعة متغيرات (نسبة السود، نسبة البيض، نسبة الأسيويين، وهكذا).

وسيكون لهذا البعد العرقي ترجيحاً كبيراً في تحديد التجميعات مثلما يقوم به عدد المتغيرات المستخدمة، ولأنها ممثلة بثلاث متغيرات أو أكثر قد تغمر قياسات أخرى (مثل حجم الساكنة التي يمكن التقاطها بمتغير واحد فقط).

ثمة عدد من الأعداد الفرعية المختلفة لتحليل التجميع، ولكننا سنركز هنا على أربعة منها متاحة في الغامب (JMP):

- التجميع التراتبي (Hierarchical Clustering).
- تجميع معدل- k (k-Means Clustering).
- المزيجات العادلة (Normal Mixtures).
- خرائط التنظيم الذاتي (Self Organizing Maps).

ولكل تحليل من هذه التحليلات نقاط قوة ونقاط ضعف، التي سنناقشها بعد حين.

التجميع التراتبي

في هذه الطريقة، نبدأ حالاتنا جميعها بشكل منفصل وفردي - فكر في كُلّ حالة باعتبارها تجميعاً صغيراً لها، عضو مستقل. ومن أصل هذه التجميعات، نجد التجميعين الأكثر قرباً، ونقوم بتجميعهما داخل تجميع أكبر. ونكرر عملية ربط التجميعين الأكثر قرباً، في كُلّ خطوة، إلى أن نحصل - في النهاية - على تجميع واحد وكبير، يحتوي على كُلّ الحالات داخله. ومن ثم، فإن التجميع التراتبي هو إجراء تكتلي (Agglomerative)، يولّد خلال هذه العملية أي عدد ممكن من التجميعات بين واحد وعدد الحالات في البيانات. وتكون التجميعات الصغرى المحصل عليها سابقاً في العملية - في بعض الحالات - متداخلة فيما بينها داخل تجميعات أكبر، سَتُشَكَّل لاحقاً، وسيتم توضيح كيفية تجميع الحالات بعد المعلومة في رسم بياني معروف باسم الرسم البياني الشجري (Dendrogram).

لقد ناقشنا سابقاً كيفية تحديد مدى اقتراب حالتين أو ترصدين فردين أو تماثلهما، ولكن التجميع التراتبي عادة ما يربط ليس فقط حالتين وإنما تجميعين، بحيث يحتوي كُلّ تجمع على حالات متعددة. فكيف تحدد المسافة بين تجميعين؟ هناك أربع طرق لحساب ذلك في «الغامب» بحيث (يجب اختيار واحد منها من قبل المستخدم منذ البداية).

- يُعرِّف التجميع ذو الربط الواحد (Single-Linkage)، المسافة بين تجميعين باعتبارها الحدّ الأدنى للمسافة بين أي عضو من أعضاء التجميع الأول وأي عضو من أعضاء التجميع الثاني.
- وفي المقابل يعرف التجميع ذو الربط الكامل (Complet-Linkage)، المسافة باعتبارها الحدّ الأقصى للمسافة بين أي من العضوين من أعضاء هذين التجميعين. وتعد هذان الطريقتان لتحديد المسافة حساسة بشكل كبير لحالات النشاز (Outliers).
- كما يعد التوافق الحاصل بينهما تجميع متوسط الربط (Average-Linkage) الذي يستخدم متوسط المسافة بين كُلّ أعضاء التجميعين.

• أما تجميع ربط الجناح (Ward Linkage)، فهو أكثر تعقيداً، ذلك بأنه يمزج التجميعين الذين سينتج اتّحادهما أصغر نمو إجمالًا داخل تباين التجميع، كما تم تحديد ذلك من قبل دالة ما (عادة مجموع خطأ المربعات).

وفي آخر التجميع التراتبي، وكما تم الإشارة إلى ذلك آنفاً، لن يكون لدينا عدد من المجموعات المنفصلة، وإنما كتلة كبيرة من الحالات المترابطة معاً بشكل تراكمي، ولكن تتجلى فكرة التجميع في خلق مجموعات متميزة. كيف يمكن لنا تفسير الكتلة الكبيرة من الحالات إلى تجميعات منفصلة كنا بصدد البحث عنها؟ وكيف يحدد عدد المجموعات التي يجب أن تكون هنالك؟

إنَّ الجواب عن السؤال الثاني سيساعدنا على الإجابة عن السؤال الأول. وتذكر أن التجميع التراتبي يولّد أي عدد من التجميعات بين 1 وn، بحيث إن الأخير عدد الحالات في بياناتنا. وفي نهاية المطاف، يتوقف علينا البث في عدد التجميعات الواجب حيازتها. ولكن لدينا دليلاً نستنير به في هذا القرار، من خلال مراقبة الرسم البياني الشجري وتاريخ التجميع. وبعد تشغيل روتين تجمع تراتبي، سيتم إنتاج رسم البياني الشجري في «الغامب». وإذا حوّلْتَ مقياس الرسم البياني الشجري إلى مقياس المسافة (مثلث أحمر مقياس الرسم البياني الشجري مقياس المسافة)، فسيوضح هذا مقدار المسافة النسبية التي تم عبورها لربط تجميعين. وفي ظل ذلك، سيُولَّد رسماً بيانياً (Plot) ركامياً مستطيلاً، والذي سيرسم بيانياً النظام التسلسلي للتجميع من خلال المسافة بين التجميعات المترابطة. وفي كُلّ من الرسم البياني الركامي والرسم البياني الشجري، نسعى إلى تحديد « نقطة فاصلة طبيعية» التي في حدودها تزداد المسافة بين التجميعات بشكل سريع (وهذا شبيه باستخدام رسم بياني ركامي لتحديد عدد العوامل المستخدمة في التحليل العاملي (Factor Analysis). ويمكن القيام بهذا أيضاً رقمياً بفحص تاريخ التجميع.

إن جوابنا عن السؤالـ «كم عدد التجميعات»؟ يجيب بدوره عن السؤالـ «ما هي الحالات التي تدخل ضمن كُلّ تجميع؟». ولأن الحالات ترتبط ارتباطاً تسلسلياً حسب المسافة التي تفصلها، فإننا - ببساطة، ومن خلال اختيار عدد التجميعات - نبت في المكان المثالي الذي تقف فيه عملية التكتل. وستكون الحالات في أي تجميع كانت، قد حلت به في هذه المرحلة.

يُوصَى بالتجميع التراتبي أساساً بالنسبة إلى مجموعات البيانات الصغيرة، ذات 200 حالة أو أقل من ذلك. وفي الحقيقة، هذا مثالي بالنسبة إلى بيانات من هذا الحجم، بما أنها أقل حساسية لتأثير حالات النشاز في مجموعات البيانات الصغيرة، مقارنة بطرق أخرى ستناقش لاحقاً، خاصة تجميع معدل k-. وبهذه الأعداد الهائلة من الحالات، يميل التجميع التراتبي إلى أن يكون مكثفاً حاسوبياً، وتُفضّل طرق أخرى.

التجميع التراتبي في «الغامب»

سنقوم باستخدام التجميع التراتبي لتجميع محافظات الولايات المتحدة في مجموعة بيانات انتخابات عام 2012. والآن، لاستخدام التجميع التراتبي في الإطار الأنسب، قمنا بانتقاء – بشكل عشوائي – فقط عدد صغير من المحافظات (75، تحديداً).

نقوم بفتح مربع الحوار (Dialog) للتجميع: (Analyze > Multivariate Methods > Cluster).

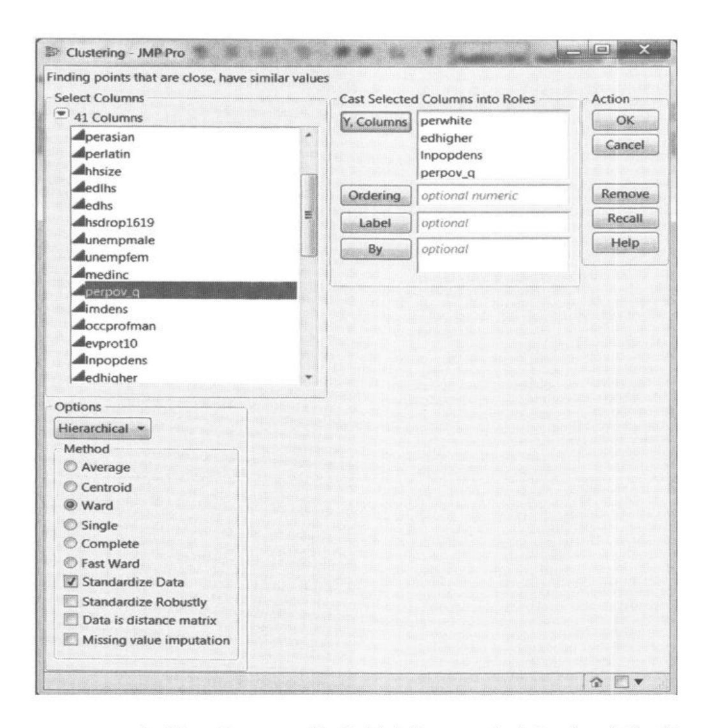
وفي هذه النوافذ (الشكل رقم 1.12)، يحدد التجميع التراتبي باعتباره الافتراض (Default) ضمن خيارات (Options) في أسفل اليسار. ويسمح البرنامج باختيار دلالات (Functions) الربط، وننتقي الجناح (Ward). أما خيار البيانات الموحدة أو المعقدة (Data Standardize)، فيتم التحقق منه، وهذه سمة لطيفة، لأننا نريد أن تكون المُدخلات على مقياس واحد.

وننتقي كمتغيرات مُدْخل، نسبة المحافظة التي تُعرّفُ بغير البيض المنحدرين من الأسبان، ونسبة الحاملين لدرجة الباكالوريوس أو درجة أكبر، ومعدل الفقر، والخوارزمية الطبيعية للكثافة السكانية، ونقرنا فوق (OK)، يعطي عملية الانطلاق للتحليل.

ويتم إنتاج الرسم البياني الشجري والرسم البياني الركامي، وتاريخ التجميع، بشكل تلقائي، وسنستخدم هذه الرسومات البيانية (Charts)، بالإضافة إلى تاريخ التجميع، لاختيار تجميعاتنا. ونقوم بتكييف الرسم البياني الشجري ليعكس المسافات (المثلث الأحمر > الرسم البياني الشجري > المقياس > المسافة) (Red

Triangle > Dendrogram > Scale > Distance)، ولجعل التجميعات متميزة بصرياً (مثلث أحمر > تجميعات اللون > مثلث أحمر > تجميعات العلامة) (Red Triangle > Color Clusters; Red Triangle > Mark Clusters).

أما الرسم البياني الشجري والرسم البياني الركامي فهما مبينان في الشكل رقم 2.12.

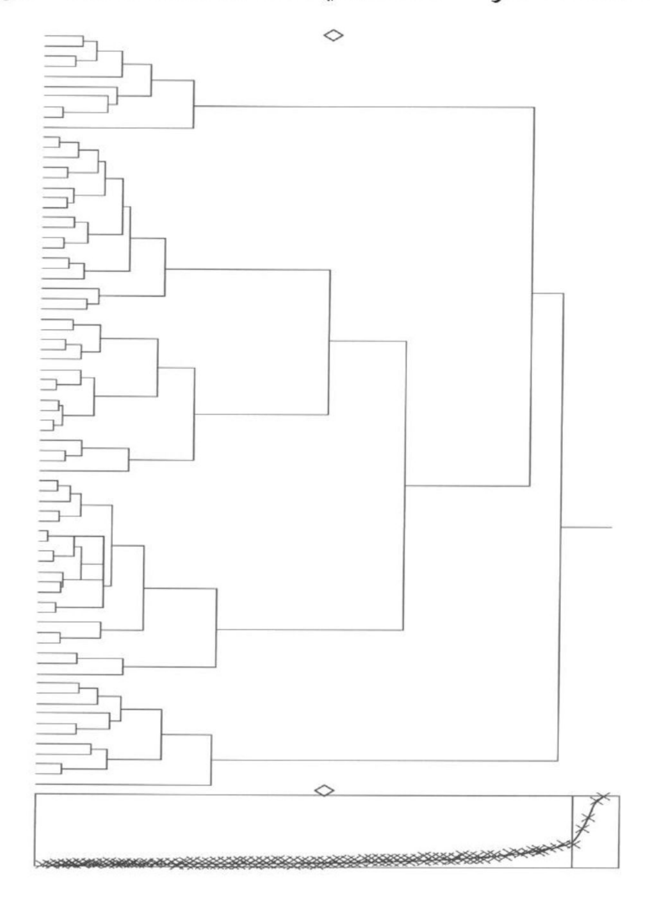


الشكل رقم 1.12: منصة إطلاق التجميع في «الغامب برو».

وسيساعدنا هذا التصور على اختيار عدد التجميعات التي سنحتفظ بها. وثمة علامة صغيرة في شكل مَعِين (Diamond-Shaped) في أعلى الرسم البياني الشجري وأسفلها. وبنقل هذا يميناً ويساراً، يكون من الممكن تغيير عدد التجميعات. كما يمكن أيضاً البحث عن الرسم البياني الركامي - عن مكان بدأت فيه المسافة بين التجميعات في الارتفاع بشكل شديد الانحدار - نحو النهاية، أي خمس حالات مزج من أقصى اليمين. وهذا عدد جيد لتجميعات نهائية. وللتحقق من أن هذا اختياراً

جيداً، يمكن النظر إلى تاريخ التجميع، الذي يظهر في «الغامب» تحت الرسم البياني الشجري (ولكن غير مبين هنا). ونبحث عن نقطة تبدأ فيها المسافة بين التجميعات في الارتفاع بشكل أسرع من ذي قبل. وهنا يتطلب المرور من خمس تجميعات إلى أربعة عبور مسافة 85.61 في حين يتطلب المرور من ستة تجميعات إلى خمسة عبور مسافة 0.07 فقط ومن ثم تستقر عند خمس تجمعات كحل جيد.

وبمجرد الحصول على عدد التجميعات التي تريدها، يكون من الممكن حفظ التجميعات (مثلث أحمر > حفظ التجميعات) (Red Triangle < Save Clusters).



الشكل رقم 2.12: رسم بياني شجري يصف تجميع الحالات المستخصلة من روتين التجميع الترابي «للغامب برو».

ويستنتج هذا متغيراً جديداً في البيانات يدعى «التجمع» وقد تريد أيضاً حفظ ترتيب العرض (Save Display Order) الذي يعمل على حفظ ترتيب الحالات في الرسم البياني الشجري القائمة من الأعلى إلى الأسفل. ويمكنك بعد ذلك استكشاف مدى اختلاف المتغيرات على مستوى التجميع، الذي يبين معنى التجميعات (الجدول رقم 1.12). وفي هذه البيانات ومن بين عيناتنا التي تصل إلى 75 حالة، يكون لتجميع 1، النسبة الأكثر انخفاضاً للسكان المكوّنة من البيض غير الإسبان، وأعلى معدل الفقر، والكثافة السكانية الأكثر انخفاضاً. أما تجميع 2، فلديه نسبة عالية نسبياً من البيض، والكثافة السكانية، ومتوسط المعدلات الخاصة بالفقر والبالغين ممن لهم تعليماً جامعياً. وأما تجميع 3، فله أدنى نسبة من السكان، ممن بحوزتهم شهادة جامعية، ومعدل فقر عالي نسبياً وكثافة عالية - إلى حدّ ما - من البيض. ولتجميع 4 أكبر نسبة من البيض، وذات كثافة سكانية منخفضة للغاية. وأخيراً يضم تجميع 5، محافظات لديها في المتوسط، ساكنة بحوزة بالغيها تعليماً عالياً وكثافة سكانية عالية، ومعدل فقر منخفض.

	تجميع 1	تجميع 2	تجميع 3	تجميع 4	تجميع 5
أبيض ٪	52.05	85.65	88.47	91.22	79.81
تعليم عالي ٪	16.85	21.27	11.95	18.53	35.97
معدل الفقر	20.45	11.21	17.12	9.15	6.58
الكثافة (log) السكانية	2.97	4.85	3.76	2.01	5.66
أوباما ٪	41.78	38.32	31.62	29.22	46.71
N	10	18	16	20	11

الجدول رقم 1.12: خصائص التجميعات المنتجة بواسطة التجميع التراتبي.

وأما معدل التصويت لصالح أوباما، فكان الأعلى في تجميعي 1 و5، مما يوافق النتائج التي عرضناها سابقاً، التي مفادها أن حصة التصويت لدى الديمقراطيين في

المحافظة، تميل إلى تكون أكبر من المحافظات المتنوعة عرقياً، وفي المحافظات ذات مستويات تعليم عالي. ومع ذلك، تذكر أن المحافظات التي قمنا بضمها هنا تمثل مجموعة فرعية عشوائية صغيرة (حوالي 2.5٪ عينة)، ومن ثم فإن التعميم الذي يقوم على أساس هذه الاستنتاجات يجب أن تتناول بحذر. وفي القسم الموالي، سنستخدم تقنيات تسمح بضم كُلّ المحافظات.

تجميع معدل k-

يختلف تجميع معدل k- الإجراء الأكثر شيوعاً – نوعاً ما – عن التجميع التراتبي. والأهم من ذلك، عدم تداخل التجميعات في معدل k-، أي إن التجميعات الكبرى لا تضمن التجميعات الصغرى بأي حال من الأحوال. وعلى العكس من ذلك، ينتج تجميع معدل k- (وهذا في الواقع يصدق على الشكلين الآخرين من التجميع الذي سنناقشه) عدداً معيناً من تجميعات مميزة (Discrete)، وذلك بتقسيم البيانات إلى أجزاء متقطعة عوض جمعها كتلة. وإن عدد التجميعات غير محدد باعتباره نتيجة لعملية التجميع، ولكن لا بُدَّ من تحديده من قبل الباحث مقدماً. وأخيراً، من الأرجح أن يجد تجميع معدل k- أكثر من التجميع التراتبي، حلولاً أقل مثالية، تحتاج نوعاً ما، إلى عمل رقابي من لدن الباحث.

وفي بداية تجميع معدل k-، يحدد الباحث k الذي يشير إلى عدد التجميعات التي ينبغي إيجادها في البيانات إلى جانب مجموعة متغيرات المدخل. ويستمر البرنامج في اختيار نقاط k- بشكل عشوائي في حيز متعدد المتغيرات. (وفي أغلب الأحيان، تقوم بهذا، من خلال اختيار مجموعة نقاط أو حالات البيانات الحقيقية. وتصبح هذه النقاط مراكز (أو «النقاط الوسطى») (Centroids) للتجميعات. وبعد ذلك يحسب تجميع معدل k-، المسافة (الإقليدية) بين كُلّ حالة، وكل نقطة من النقاط الوسطى و«تخصيص» حالة النقاط الوسطى الأقرب. ونحصل من ثم على نقاط وسطى k بحيث يحيط بها «سحابة» مشوهة من الحالات وبعد ذلك يجد تجميع معدل k–، المعدل أو المركز لكُلّ سحابة من سحابات النقطة (ومن غير المرجح أن يكون المعدل هو النقطة المختارة في البداية) ويجعل هذه النقاط، نقاط المركز الجديد. وتعيد الخطوات نفسها التي سبقت – وذلك بحساب المسافات، وتخصيص

حالات للنقاط الوسطى، وإيجاد نقاط المعدل، وتحويل النقاط الوسطى، مراراً وتكراراً إلى أن يقارب البرنامج حلاً مستقراً. وفي هذه النقطة، لدينا مجموعة من تجميعات k ، بحيث يتكون كُلّ واحد منها من عدد معين من الحالات.

والسؤال البديهي الذي يرجى حله، هو كيفية الشروع في اختيار عدد التجميعات التي نريدها. هناك جوابان ممكنان عن هذا السؤال. يمكننا اختيار قيمتنا لـ k وفق نظرية من النظريات. قد نختار ثلاثة أنظمة من أنظمة الرعاية الاجتماعية للدولة – مثلاً – إذا استرشدنا بتصنيف (Typology) إسبينغ- أنديرسون (Esping-Andersen). ومع ذلك، قد يكون هذا، أو قد لا يكون عدد التجميعات المثالية المحصل عليها بشكل تجريبي، مما يحيلنا على المقارنة الثانية. وفي هذا الحل، نستمر في عملنا مثل مختص حقيقي في التنقيب في البيانات، ونجرب عدداً من القيم المختلفة لـ k (عادة على نطاق معين)، وانتقاء القيمة التي يكون فيها الحلّ الأفضل.

ولكن كيف يتسنى لنا معرفة الحلّ «الأفضل»؟ في الواقع، ثمة نوعان من إحصاءات التناسب، يمكن الاستعانة بهما لتحديد ذلك. أما نوع الإحصاء الأكثر إفادة في هذه الحالة، فهو نسبة التباين (Dissimilarity Ratio)؛ نسبة المسافة بين التجميعات إلى نسبة المسافة داخل التجميعات. وعلينا اختيار عدد التجميعات التي تعظم هذه النسبة، والشيء المثير بشأن هذا القياس، هو أن نسبة التباين - وبخلاف قياسات تناسب أخرى (مثل مجموع أخطاء المربع) - لا تنخفض آلياً لدى إضافتنا التجميعات. وإن إضافة التجميعات قد يقلص مسافة التجميعات من الداخل (ويعني مزيد من التجميعات، أن كُلّ تجميع سيشغل حيزاً أصغر، ويضم حالات أقل، ولكن قد تقلص أيضاً مسافة التجميع البيني (ويعني مزيد من التجميعات في الحيز المتعدد الأبعاد نفسه، أن التجميعات نفسها معبأة بإحكام أكثر). وبالتالي، من المرجح أن يكون حلّ «مثالي» للسؤال الخاص بعدد التجميعات الواجب تحديديها باستخدام نسبة التباين. ولكن، لسوء الحظ، لا ينتج «الغامب» هذه النسبة آلياً (كما يجب). فعلى المستخدمين حسابه بأنفسهم. أما الطريقة المتبعة في القيام بذلك، فسيتم وصفها لاحقاً.

يساعدنا استخدام نسبة التباين على اختيار k، ولكن هذا لا يضمن لنا – على الإطلاق – إيجاد حلّ مثالي. ولفهم هذا، تذكر كيفية اختيار نقاط التجميع الأولى:

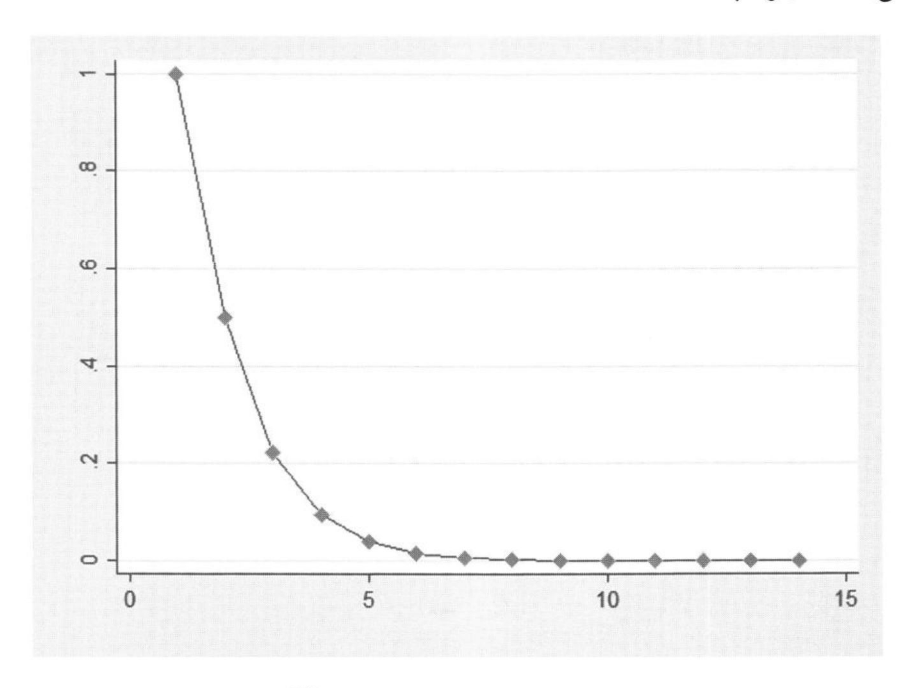
عشوائياً. إن عملية دعم العشوائية يساعد على إزالة التحيز الذاتي من الإجراء، ولكن لديها الجودة غير المناسبة، المتمثلة في ندرة إيجادها الحلّ الأنسب لمشكل ما. كما تساعد عملية التكرار عبر الخوارزمية لنقل نقطة المركز - لا محالة - على تصحيح ذلك إلى حدّ ما، ولكن يبقى الحلّ النهائي - مع الأسف - متأثراً بالقيم المختارة عشوائياً في البداية.

ولندرس ما يلي: ودعنا نقول إن بياناتنا تحتوي على تجميعات k «حقيقة». ودعنا نقول إننا محظوظون، ونختار القيمة نفسها لـ k بالنسبة إلى عدد التجميعات الموجودة مع برنامج تجميعنا لمعدل k- ويختار البرنامج النقاط المركزية لـ kبشكل عشوائي ويشرع في التكرار. ولكن، احتمال سماح العشوائية لنا باختيار النقاط المركزية، بحيث يكون لكُلّ تجميع «حقيقي» نقطة مركزية واحدة، وواحدة فقط داخلها، مختارة على نحو منخفض جداً؛ إنه بالأحرى مثل معرفة وجود ثمانية أهداف دائرية (المستخدمين في الرشق بالسهام) (Dart ،Boards) على حائط ما؛ فإذا رمينا ثمانٍ سهام عشوائياً على الحائط، فستستقبل - من الأرجح - بعض الأهداف الدائرية سهام متعددة داخلها، في حين لا تستقبل أخرى، أي سهم.

وإذا كان لكُلّ التجميعات الحقيقية الحجم نفسه (الذي يعد الأفضل بالنسبة إلى غاياتنا، فإن عدد الطرق التي قد نختار من خلالها نقطة واحدة لكُلّ تجميع هو !k ($1\times2\times3\times\ldots\times k$)، ولكن عدد الطرق التي نستطيع من خلالها اختيار نقاط k، هي k^k (مع افتراض أن حيزنا المتعدد المتغيرات كله هو في منطقة تجميع من التجميعات). وهذا يعني - عموماً - أن احتمال اختيارنا لنقطة واحدة لكُلّ تجميع ($P= k! /k^k$) منخفض، ويهبط أكثر عندما ترتفع عدد التجميعات (انظر الشكل رقم 3.12). وبمجرد أن تكون لدينا خمس تجميعات، يهبط احتمال الاختيار الأولي لنقطة واحدة من النقاط الوسطى لكُلّ تجميع، إلى 0.038 وعندما تكون لدينا 11 تجميعاً، تكون لدينا احتمالات تصل إلى حوالي واحدة في 10.000.

ولقد تمت الإشارة إلى ذلك سابقاً، إن نقل النقط الوسطى بشكل متكرر، يساعدنا - إلى حدّ ما - ولكن لا يضمن الالتقاء عند حلّ مثالي. وتكون الاحتمالات أكثر انخفاضاً إذا لم تكن التجميعات متساوية في الحجم، أو في الكثافة، أو في «الشكل الكروي» - وستكون الاحتمالات في البيانات الحقيقية - على الأقل ستتم مصادفة

مشكلة من هذه المشاكل. وستتم طرق معالجة هذه المشكلة، ذات تجميع معدل k- من ثلاث جوانب.



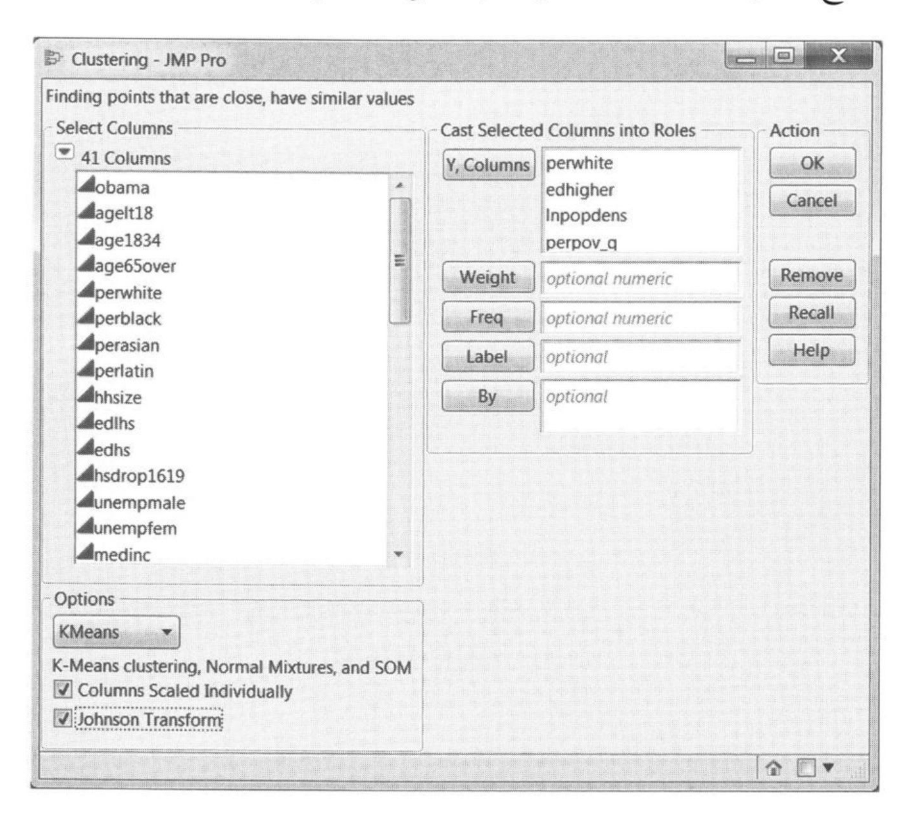
الشكل رقم 3.12: احتمال انتقاء نقطة وسطى واحدة لكُلّ تجميع «حقيقي» بواسطة عدد التجميعات «الحقيقية».

أولاً: نستطيع داخل أي قيمة معينة لـ k، القيام بمحاولة تقليص مجموع أخطاء المربع، وتعظيم نسبة التباين. ويشمل ذلك، إعادة تشغيل الخوارزمية عدة مرات، سعياً إلى البحث عن الحلّ «الأفضل»، لكن أيضاً سعياً إلى البحث عن الحلول التي تتكرر - بحيث تنتهي الحالات نفسها في تجميع واحد مراراً وتكراراً.

ثانياً: يمكننا استخدام طرق التصور، بالبحث عن الحالات النائية عن المركز التي قد تنحرف عن النتائج، ومراقبة التجميعات نفسها لفحص ما، إن كان الحلّ ممكناً (يسمح «الغامب» بالنظر إلى حلّ التجميع ذي الرسوم البيانية لـ مكون رئيسي ثنائي أو ثلاثي الأبعاد).

ثالثاً: يجب علينا تذكر أن وجود أي عدد من التجميعات «الحقيقية» في بياناتنا،

هو أمر غير مرجح تماماً. ومن الأفضل اعتبار حلول التجميع بمثابة استدلال (Heuristics)، نسترشد بها لتبسيط البيانات والبحث عن الأنماط المهمة، عوض الكشف عن الطبقات الخفية للحقيقة. ومن ثم، إذا صح ذلك، فإن نسبة «صحة» حلّ التجميع سيكون نسبياً دائماً. وعلى نحو مماثل، قد تكون هذه النسبة «خاطئة».



الشكل رقم 4.12: اختيار تجميع معدل k- في منصة إطلاق التجميع «للغامب برو».

تجميع معدل k- في «الغامب»

سنستخدم بيانات انتخابات 2012 على مستوى المحافظة (التي استخدمناها في التجميعات التراتبية) لإنجاز تجميع تجميع معدل k-، ولكن سنستخدم هذه المرة 3114 محافظة برمتها، عوض استخدام عينة صغيرة منها. ويتم ذلك من خلال فتح البيانات، وإيجاد العلبة الأولى لحوار التجميع (Clustering Dialog Box) (الشكل

رقم 4.12). وفي القائمة المندرجة تحت الخيارات في الزاوية السفلية اليسرى، نغير الإعداد من التراتبية (Hierarchical) إلى معدل K (K–Means). ونقوم أيضاً بتحويل متغيرات المدخل عبر اختيار تحوّل جونسون (Johnson Transform)، مما يطبّع المتغيرات المنحرفة، ويكبح جماح الحالات البعيدة عن المركز. ونقوم باستخدام مجموعة متغيرات المدخل نفسها التي استخدمناها بالنسبة إلى التجميع التراتبي، ولكن نضيف حصة أوباما من التصويت، ونسبة الساكنة السوداء، وذات الدخل المتوسط. وتفتح منصة إطلاق التجميع التكراري (Iterative Clustering Launch Platform) (وللاطّلاع على ظهوره العام، انظر الشكل رقم 7.12 في الفصل الموالي).

نتتقي عدداً من التجميعات - أو عوض ذلك - مسافة بالنسبة إلى K. ونمكّن البرنامج من منحنا نتائج بالنسبة إلى التجميع 3 والتجميع 5 (غير مبينة). كما أنه أيضاً فكرة جيدة لاستخدام انحرافات معيارية داخل التجميع، لأن ذلك سيساعد على حساب الإحصاءات التناسبية لاحقاً.

لقد استخدمنا الساكنة بأكملها لـ 3,441 محافظة، ونرى في الشكل رقم 5.12 أن معظمها انتهى بتجميع واحد (تجميع 2). ولدى العديد من التجميعات المولّدة الأخرى أعداداً صغيرة من الحالات. وقد يعني هذا:

1. أن بياناتنا غير قابلة للتجميع، أو
2. أننا اخترنا العدد الخاطئ للتجميعات، أو
3. أننا وجدنا «حلاً محلياً» غير مثالي، أو
4. أن بياناتنا الحقيقية تتألف من مجموعة كبيرة من حالات مماثلة ذات مجموعات متباينة استثنائية.

ويمكننا التحقق من هذا من خلال إعادة إجراء التحليل. ولكن لاحظ أنه إذا ما قمنا ببساطة، «بإعادة إطلاق التحليل» (Relaunch Analysis)، فستُخدم قيم البذور نفسها، وسنحصل على حلّ متطابق. إننا في حاجة إلى البدء من الصفر للحصول حلّ تجميع مختلف.

ولإنتاج مجموع أخطاء المربع بالنسبة إلى النموذج، ننقر المثلث الأحمر

المجاور لـ"K Means NCluster =3"، ونختار حفّظ التجميعات (Save Clusters). وسينتج هذا عمودين جديدين: مهمة التجميع، وعمود يدعى المسافة (Distance)، التي تعد مسافة كُلّ حالة مستقلة من نقطتها الوسطى (Centroid). وننتج عموداً ثالثاً الذي يقوم بتربيع هذه المسافات. وبعدها، نحسب معدل متغير مربع المسافة، ونضربه في عدد الحالات في التحليل. وهذا هو مجموع الأخطاء المربعة.

K Means NCluster=3

Columns Scaled Individually, Use within-cluster std deviations

Cluster Summary

Cluster	Count
1	136
2	2890
3	88

Step	Criterion
24	0

Cluster Means

Cluster	perwhite	edhigher	lnpopdens	perpov_q	obama	medinc	perblack
1	0.97505188	0.24401959	-1.7185925	-0.6253437	-1.2323663	-0.1389045	-1.7367295
2	-0.0868893	0.03458119	0.08513217	-0.0041018	0.07714976	0.05269747	0.10880101
3	1.34662458	-1.5127989	-0.1398114	1.10114552	-0.6291022	-1.515962	-0.8890874

الشكل رقم 5.12: مُخرج معدل k- في «الغامب برو».

ولإنتاج نسبة التباين (الجدول رقم 2.12)، نحصل في، أولاً على إحصائيات وصفية أساسية (المعدل والانحراف المعياري) لكُلّ متغير مدخل على حدة. ومن خلال استخدام هذه الإحصاءات الوصفية - بعد ذلك - إلى جانب النتائج المدرجة تحت «المقياس الأصلي لمراكز التجميع» (Cluster Centers Original Scale) لكُلّ متغير، نحسب نتيجة -z لمركز التجميع. وسيكون هذا مختلف عن معدل تجميع «تحوّل جونسون». ونحسب المسافة الإقليدية بين كُلّ مجموعة من مجموعات نتيجة -z لمراكز التجميع، واتّخاذ الأصغر من أصل هذه المسافات باعتباره قياس المسافة بين التجميع. وبعد ذلك، نجد لكُلّ تجميع، المسافة القصوى

لحالة ما إلى النقطة الوسطى، وأخذ متوسط هذه المسافات القصوى باعتبارها قياسنا للمسافة داخل التجميع. وللحصول على نسبة التباين، نقسم مسافة بين التجميع على مسافة داخل التجميع.

ويفترض هذا التحليل، تفوق حلّ التجميع الرابع قليلًا، على حلّ التجميع الثالث والتجميع الخامس. وإن الأرقام العالية بالنسبة إلى متوسط الحدّ الأقصى لمسافة داخل التجميع، يمكن أن تتأثر بحضور الحالات الشاذة في البيانات. إن معدل المسافات التجميعات أصغر بكثير. ونستطيع استكشاف هذه الإمكانية من خلال فحص رسم بياني ثنائي، ثلاثي الأبعاد. وسيوضح هذا كيف أن الحالات والتجميعات منظمة في الحيز الثلاثي الأبعاد، والمحددة بالمكوّنات الأساسية الثلاثة الأولى لمتغيرات المدخل. ويمكن توليد هذا من خلال فتح القائمة بجانب حلّ التجميع الذي نهتم به في فحص واختيار الرسم البياني الثنائي الثلاثي الأبعاد (Biplot 3D). ونعرض الرسم البياني الثنائي، الثلاثي الأبعاد لهذا التحليل في الشكل رقم 6.12. ويكشف الرسم البياني عن بنية بياناتنا. ولم تتجمع الحالات في مناطق متفرقة جداً، بل إنها مجمعة في اتّجاه مركز الحيز بشكل عام (في تجربتنا، تُعدّ هذه أكثر شيوعاً من البيانات «المجمعة» بشكل واضح). وكل تجميع أيضاً لديه حالات متعددة مخصصة له، التي تشكل حالات استثنائية بشكل واضح.

إن حلّ التجميع الرابع في هذه البيانات، تحدد محافظات مختلفة (الجدول رقم 3.12)؛ ففي:

التجميع الأول: لدينا مجموعة صغيرة من المحافظات القوقازية بشكل كبير، وفقير جداً في المتوسط. وكان لدى هذه المحافظات أدنى معدل دعم لأوباما في العام 2012 من أصل كُلّ التجميعات.

التجميع الثاني: فيشكل غالبية السكان البيض - ولكنه أقل كثافة سكانية - ذات متوسط دخل أعلى، ومعدل فقر أقل. وإن حصة أوباما من أصوات في هذه الدول كانت أعلى شيئاً ما من التجميع الأول.

التجميع الثالث: فهو التجميع النموذجي. وإنه مُتَنوّع إثنياً بقدر أكبر من كُلّ من التجميع الأول أو الثاني، ولديه معدل فقر قريب من المعدل الوطني.

التجميع الرابع: نجد فيه محافظات متنوعة إثنياً، لها العديد من طلبة الجامعة، ومتوسط دخل عالٍ، ولها كثافة سكانية عالية نسبياً (حوالي 375 شخصاً في الميل المربع).

الجدول رقم 2.12: إحصاء التناسب بالنسبة إلى تجميع معدل k-.

	مجموع الأخطاء المربعة	مسافة بين التجميع الصغرى	متوسط الحدّ الأقصى لمسافة داخل التجميع	نسبة التباين
3	253,043.64	1.52	36.13	0.041453
4	252,981.36	2.07	49.93667	0.041453
5	250,739.28	1.64	31.49333	0.052075

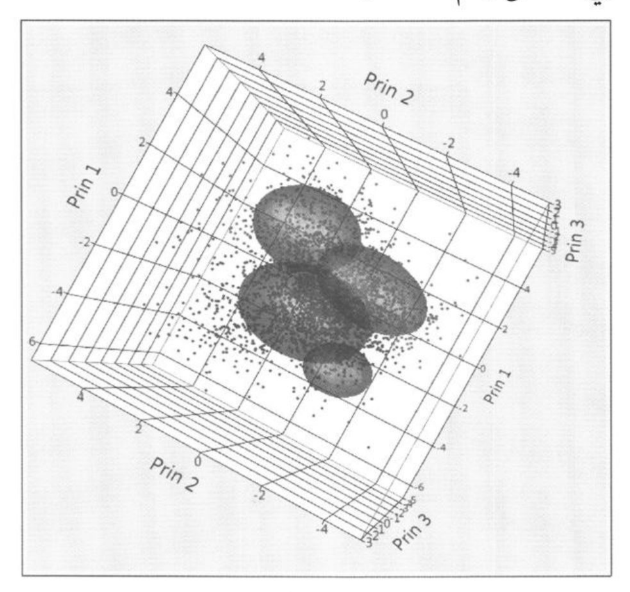
تمازجات عادية

إن التمازجات العادية (Normal Mixtures) وتمازجاتها العادية المتينة (Robust Normal Mixtures) الشقيقة شبيهة للغاية بتجميع معدل k-. ويكمن الفرق الرئيس في تخصيص الحالات للتجميعات. وفي تجميع معدل k-، يُخصص لكُلّ حالة تجميع واحد. وفي المقابل، تقوم التمازجات العادية بحساب احتمالية أن تكون حالة معينة في كُلّ تجميع. ويقوم النموذج أولاً برسم خريطة الحيز بمجموعة من التوزيعات العادية متعددة المتغيرات التي تعمل بصفتها «تجميعات». وستكون لكُلّ حالة قيمة معينة في توزيع عادي متعدد المتغيرات للتجميع. وكما هو الحال بالنسبة إلى معدل k-، فإن النقاط المركزية لهذه التجميعات تتكرر إلى حين إيجاد حلّ مستقر، ولكنه محلي، احتمالاً.

الشيء الجميل بشأن التمازجات العادية هو أنها تنتج – كجزء من مخرجها

الأساسي – مصفوفة تباين التغاير لمتغيرات المدخل بالنسبة إلى كُلّ تجميع. ويسمح لنا هذا بالبحث في كيف أن الارتباطات بين المتغيرات، تختلف بين التجميعات، ويمكن أن تساعد إذا أردنا تحويل حلّ تجميع إلى انحدار مع متغيرات تفاعل. إضافة إلى ذلك، إن مسألة اعتبار الاحتمالات تقديرات لكُلّ زوج تجميع – حالة، تسمح لنا بتحديد حالات على الشريط الحدودي لأن تصبح في تجميعات متعددة.

وفي «الغامب»، يتم أداء تجميع التمازجات العادية بشكل كبير على النحو نفسه التي ينجز به معدل –k. وبمجرد الحصول على منصة التجميع التكراري، غير ببساطة «معدل k-» إلى «تمازجات عادية» أو إلى «تمازجات عادية متينة». وستظهر المنصة كما هو الحال في الشكل رقم 7.12 أو 8.12.

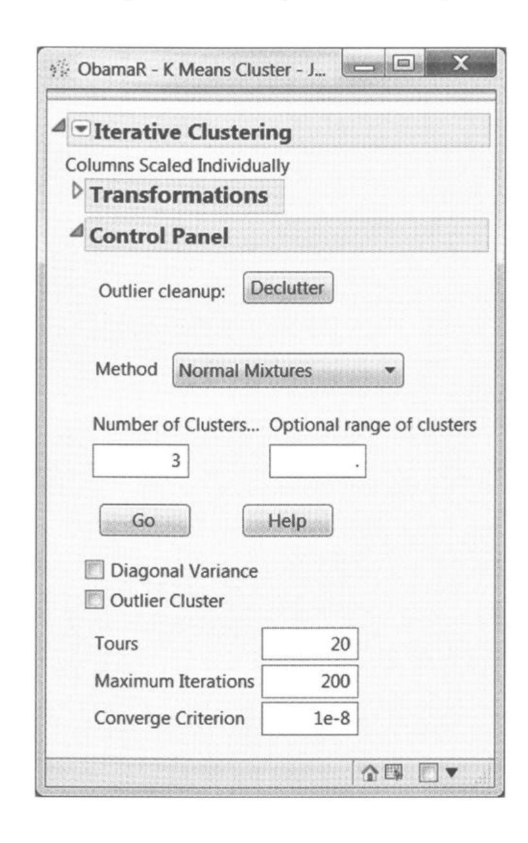


الشكل رقم 6.12: رسم بياني ثنائي، ثلاثي الأبعاد، يبين التجميعات المنتجة بواسطة تجميع معدل k-.

الجدول رقم 3.12: خصائص التجميعات المنتجة بواسطة تجميع معدل k-.

	تجميع 1	**تجميع 2**	**تجميع 3**	**تجميع 4**
أبيض ٪	97.46	92.51	74.25	74.43
تعليم عالي ٪	9.22	19.70	15.09	30.21

معدل الفقر	22.22	9.34	17.49	10.23
الكثافة (log) السكانية	3.45	2.63	3.59	5.92
أوباما ٪	28.18	33.24	37.83	47.49
أسود ٪	0.19	0.51	12.35	10.14
الدخل المتوسط	29,300	47,570	37800	59000
N	46	749	1721	598

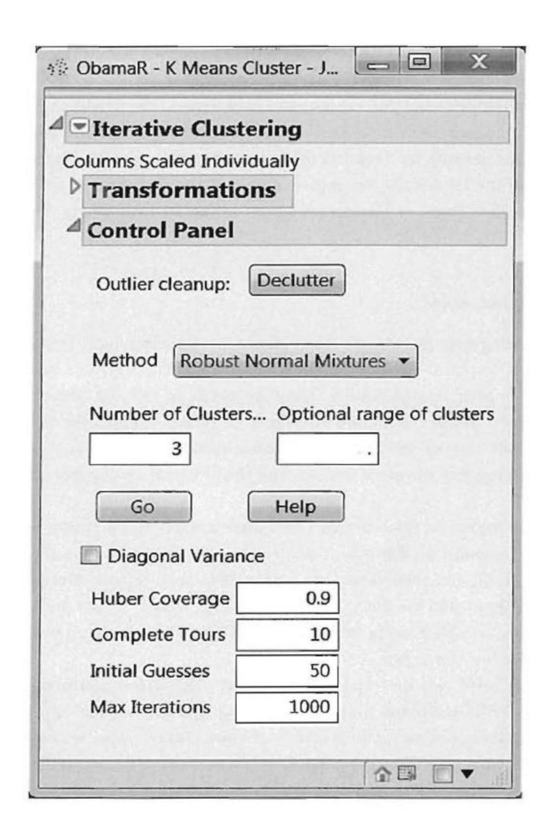


الشكل رقم 7.12: اختيار تجميع التمازجات العادية في منصة إطلاق التجميع التكراري.

وتظهر هنا بعض الفوارق انطلاقاً من تجميع معدل k-. ولاحظ أن البرنامجين كلاهما يسمحان بأعداد مختلفة من «الجولات»، من أجل التحذير من إيجاد حلّ تجميع محلي، من خلال تشغيل البرنامج بقيم بداية مختلفة (The Max) (الحدّ

الأقصى). إن إعداد التكرارات يقيد خوارزمية التهيئة (Optimization) بالنسبة إلى عدد معين من التكرارات لتسريع المعالجة.

إنَّ معيار التقارب (والمتاح في التمازجات العادية، ولكن ليس في التمازجات العادية المتينة) يطلع «الغامب» على الاختلاف في الاحتمالية الخوارزمية -Log) (Likelihood التي تأخذ فيها بعين الاعتبار النموذج المتقارب، وإيقاف عملية التكرار.



الشكل رقم 8.12: اختيار قوي لمزيج طبيعي مخلوط في تكرارية منصة الإطلاق العنقودية

ثمة اختلافات قليلة بين هذين النافذتين؛ إذ لدى التمازجات العادية المتينة إعداداً (Setting) لتغطية هوبر (Huber Coverage)، وهذا تقدير مماثل لتقديرات «ساندويتش» هوبر- الأبيض المستخدمة في الأخطاء المعيارية المتينة. ويطلع

الإعداد «الغامب برو» على نسبة الحالات التي لا يجب اعتبارها حالات شاذة (Outliers)، ومن ثم لا يجب تقليص ترجيحها. وتسمح التمازجات العادية بخلق «تجميع شاذ» إضافي، يستطيع ضبط حالات تقع خارج منطقة أي من التجميعات الموجهة للمستخدم. وسيمنع هذا الحالات الشاذة من ممارسة تأثير كبير على المكان الذي توجد فيه التجميعات.

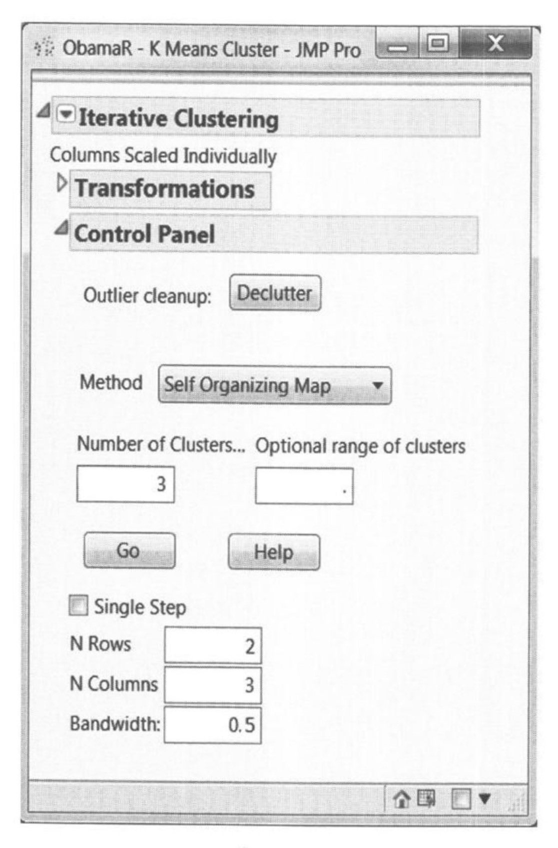
الخرائط المنظمة ذاتياً

إن معظم الخرائط المنظمة ذاتياً - مثلها مثل الشبكات العصبية - تتعلم الخوارزميات، ولكن هذا الأمر ينسحب على برنامج الخريطة المنظمة ذاتياً في «الغامب». إنه شبيه للغاية بتجميع معدل k-. إن الفائدة العامة للخرائط المنظمة ذاتياً في «الغامب»، تتمثل في قابليتها للتأويل. وقد تم تكوينها من أجل أن تظهر التجميعات في بنية ذات بعدين شبيهة بالهيكل (بحيث توافق محاورها، المكوّنين الأساسيين الأولين لمصفوفة متغير تباين التغاير). وتعد التجميعات القريبة من بعضها بعضاً أكثر تشابهاً، وأما التجميعات البعيدة عن بعضها بعضاً، فتعد أكثر تبايناً.

وإن ما يقع تحت الغطاء هو عملية رسم محور ثنائي الأبعاد باستخدام المكوّنات الأساسية الأولى، وقطع هذا الحيز إلى عدد محدد من قبل الباحث من مناطق متساوية الحجم، وقيم بذور مخصصة لكُلّ منطقة. ويستخدم معدل k- لتخصيص حالات للبذور، ويتم إيجاد المعدل لكُلّ تجميع. كما تشغل الانحدارات - بعد ذلك - لتنبؤ المعدلات، لتُفضي إلى انتقاء نقاط وسطى جديدة، ومن ثم انحدارات جديدة إلى أنه يصير للعملية نقطة التقاء.

ولبناء خريطة منظمة ذاتياً، نقوم أولًا بفتح منصة التجميع التكرارية، وتغيير «معدل k-»، إلى «خريطة منظمة ذاتياً» (الشكل رقم 9.12). وعوض اختيار عدد التجميعات، علينا - في المقابل - انتقاء عدد السطور والأعمدة التي نريدها في هيكلنا (سيكون عدد التجميعات نتيجة هذين العددين). وبعد ذلك، نضع معلم حيز النطاق الذي يؤثر في نسبة التأثير الذي تملكه تجميعات الجيران على تقديرات نقاط وسطى. ونختار بناء 3 x 2 ذي حيز نطاق أقل من 0.5.

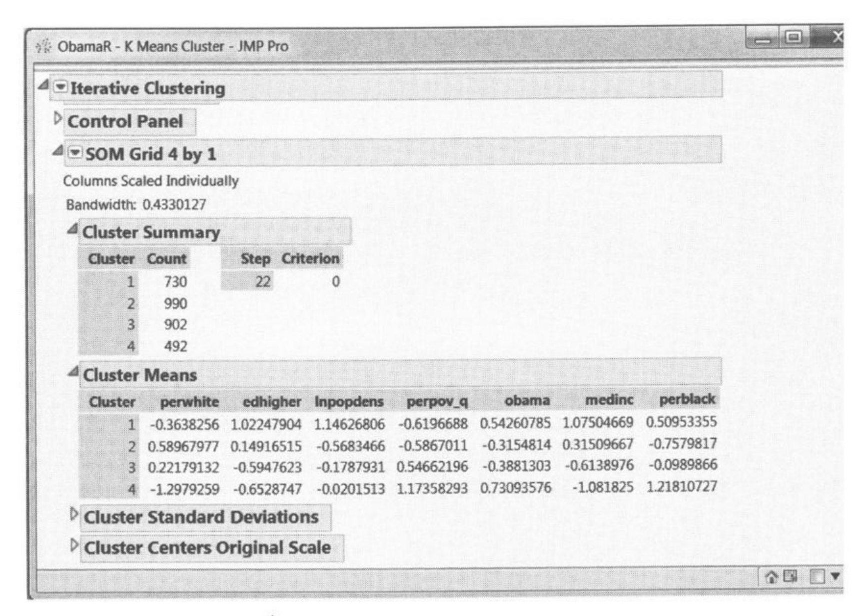
إن المطبوع الأولي شبيه جداً بمطبوع معدل k- وتمازجات عادية (الشكل رقم 10.12). ويمكن فحصه لأجل أنماط في البيانات كما هي، ولكن علينا معالجة نتائج أولية مستخلصة من خرائط منظمة ذاتياً بالقدر الذي فحصنا به نتائج مستخلصة من معدل k- وتمازجات عادية. علينا مراقبة إحصائية التناسب، ونجرب أعداداً أخرى من الحالات، وإعادة تشغيلها لتجنب حلول محلية، وهكذا.



الشكل رقم 9.12: اختيار خريطة منظمة ذاتياً في منصة إطلاق التجميع التكراري.

ويتم إعداد خريطة منظمة ذاتياً بهدف تقليصها إلى بعدين. ونعيد إنتاج الرسم البياني الثنائي بإضافة «شعاعات» المتغير (الشكل رقم 12.11). وهذا يساعد على توضيح طبيعة الخريطة المنظمة ذاتياً، ثنائية الأبعاد، ولكن أيضاً يبرز العلاقة الوطيدة بين

التجميع، وتحليل المكوّن الرئيسي. أما المكوّن الرئيسي الأول (محور أفقي) فهو مرتبط ارتباطاً وثيقاً بنسبة الساكنة ذات تعليم عالي، ودخل متوسط (بشكل إيجابي)، بالإضافة إلى معدل الفقر (بشكل سلبي). إن قياسات حصة أوباما من الأصوات، ونسبة السود، والكثافة السكانية، مترابطة بشكل كبير ومترابطة بشكل إيجابي بالمكوّن الأساسي الثاني (محور عمودي). أما نسبة البيض، فمترابطة (Correlated) بشكل سلبي بهذا المكوّن. ونرى أيضاً أن التجميعات المتنوعة تقع داخل مناطق مختلفة من الحيز المحدد من قبل المكوّنات الرئيسة. وبالتالي، إن الرسم البياني الثنائي يخبرنا بأن التجميع الأول يصف محافظات كثيفة ومتنوعة وثرية نسبياً، والثاني يصف تلك المحافظات الأكثر ثراء، ولكن أقل كثافة، وأقل تنوعاً (ومن غير المرجح أن تساند أوباما). أما التجميع الثالث، فيضم مناطق محافظات حضرية فقيرة. ويضم التجميع الرابع محافظات فقيرة، ولكنها أقل كثافة وبياضاً من التجميع 3.

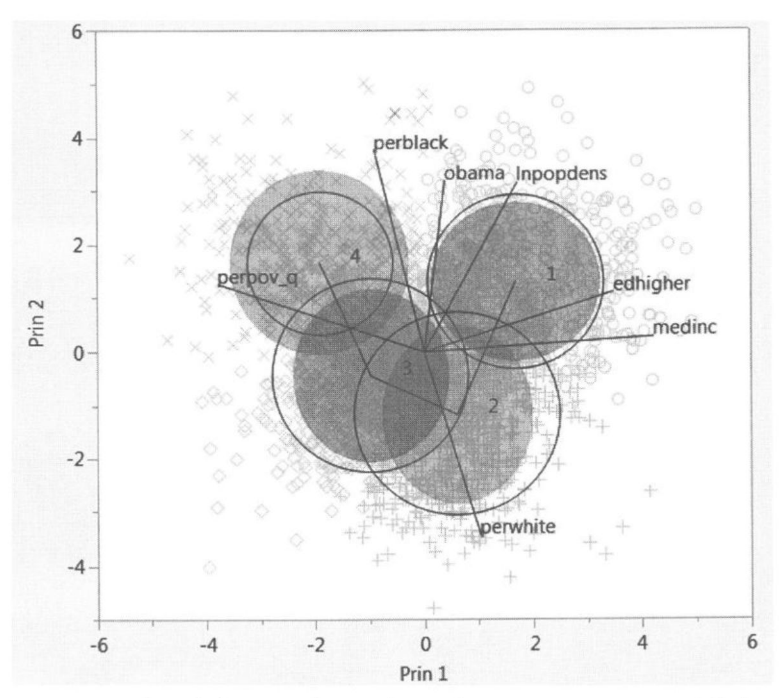


Cluster	Count
1	730
2	990
3	902
4	492

Step	Criterion
22	0

Cluster	perwhite	edhigher	lnpopdens	perpov_q	obama	medinc	perblack
1	-0.3638256	1.02247904	1.14626806	-0.6196688	0.54260785	1.07504669	0.50953355
2	0.58967977	0.14916515	-0.5683466	-0.5867011	-0.3154814	0.31509667	-0.7579817
3	0.22179132	-0.5947623	-0.1787931	0.54662196	-0.3881303	-0.6138976	-0.0989866
4	-1.2979259	-0.6528747	-0.0201513	1.17358293	0.73093576	-1.081825	1.21810727

الشكل رقم 10.12: مخرج من خريطة منظمة ذاتياً في «الغامب برو».



الشكل رقم 11.12: ثنائية الرسم البياني الواصفة لعلاقة التجميعات بالمتغيرات في خريطة منظمة ذاتياً («الغامب برو»).

وهذا يبين فائدة استخدام خرائط منظمة ذاتياً بالإضافة إلى مكونات رئيسة وشعاعات متغير لتمييز تحليلات تجميع نهائية من حيث التموضع على طول استمراريات متغيرات مترابطة في حيز متعدد الأبعاد. إن تحليلنا لأنماط التصويت لا يمثل عرضاً كاملاً، ولكن بيانات العلوم الإنسانية في تجربتنا مجمعة (Clustered) بشكل نادر وواضح. ومع ذلك، يمكن لتحليل التجميع أن يستخدم لتحديد أنماط التشابه بين حالات على مستوى مدخلات مهمة نظرياً.

الفصل الثالث عشر
تحليل الطبقة الكامنة ونماذج المزيج

تحليل الطبقة الكامنة

تم استخدام تحليل الطبقة الكامنة (Latent Class Analysis) في بداية الأمر - وبشكل بارز - في العلوم الإنسانية من قبل لازارسفيلد (Lazarsfeld) وهنري (Henry) (1968). ويعد هذا النوع من التحليل، تقنية إحصائية أخرى في العائلة الأوسع لنماذج المتغير الكامنة (Latent Variable)، إذ يضم تحليل المكوّن الأساسي (Principal Component Analysis)، وتحليل المعامل، والتجميع (Clustering).

ويمكن النظر إليه باعتباره نموذجاً، حيث يتم فيه تقدير متغير واحد كامن، له توزيع فئوي ما. وهذا الافتراض حول عدد المتغيرات الكامنة وتوزيعها، يجعل تحليل الطبقة الكامنة متميزاً عن تحليل المكوّن الأساسي، الذي يفترض وجود متغيرات كامنة متعددة ذات توزيع عادي، كما يعد تحليل الطبقة الكامنة في بعض الحالات أكثر تماثلًا من التجميع ما دام يسعى إلى استكشاف المجموعات الكامنة، ولكنها تختلف في ضرورة أن تكون متغيرات المدخل المستخدمة لإيجاد المجموعات، فئوية في حالة تحليل الطبقة الكامنة، و(في الغالب) مستمرة في حالة التجميع. ومع ذلك، يعد تحليل الطبقة الكامنة، قريباً وثيقاً للتجميع العادي للتمازجات، لأنه يفترض أن التوزيع المرصود للاستجابات مكوّن من مزيج توزيعات متعددة أكثر بساطة. وأخيراً، بما أن تحليل الطبقة الكامنة يعالج البيانات الفئوية

للمُدخل، ويقدر احتمال المتغير الفئوي الكامن، فهو أيضاً وثيق الصلة بنمذجة اللوغاريثم الخطي.

في الغالب، يُستخدم تحليل الطبقة الكامنة في تحليل البيانات الوضعية للاستجابة المستخلصة من مُسوحات (Surveys). لنتصور أننا سألنا مجموعة من الناس بشأن موافقتهم على الصلاة في المدرسة، والإجهاض، وزواج المثليين. ويمنحنا هذا مجموعة مؤلفة من ثلاث متغيرات، بحيث يأخذ كُلّ متغير قيمتين ممكنتين، فنحصل على ثمانية أنماط استجابة محتملة. نريد تصنيف الناس إلى طبقات استناداً إلى هذه الأنماط من الاستجابة، غير أننا نظن أن ثمان طبقات، كثيرة جداً. ومن خلال تحليل الطبقة الكامنة، نصنف أنماط استجابة إلى عدد أصغر من الطبقات الكامنة، محددين ذلك العدد في وقت مبكر. ويسمح لنا هذا بتقدير مجموعتين من المَعْلمات.

أولاً: نقدر انتشار كُلّ طبقة من الطبقات الكامنة.

ثانياً: نقدر احتمالية استجابة معينة لعضوية ما في طبقة كامنة. وبمثالنا الموقفي هذا، يمكن افتراض وجود مجموعتين - «الليبراليين والاجتماعيين»، و«المحافظين الاجتماعيين». واستناداً إلى بيانات استجابتنا، يمكننا تقدير نسب الأفراد الليبراليين اجتماعياً، مقابل أولئك المحافظين اجتماعياً، كما يمكننا تقدير مدى مساندة المرء، مثلاً، لزواج المثليين، باعتباره ليبرالياً اجتماعياً.

ومن الأهمية التأكيد - مع ذلك - على أن تحليل الطبقة الكامنة هو تقنية غير خاضعة للرقابة. ويشترط الباحث عدد الطبقات التي يقدرها النموذج، غير أن الحلّ الذي سيتم التوصل إليه لا يمكن تحديده من الوهلة الأولى. ومن ثم، لا نضمن، في مثالنا أعلاه - وجود مجموعات مطابقة لتصوراتنا بخصوص الليبراليين الاجتماعيين، والمحافظين الاجتماعيين. وعوضاً عن ذلك، وكما هو الحال بالنسبة إلى تحليل العامل يبقى الباحث هو المسؤول عن تأويل دلالة المجموعات الكامنة استناداً إلى توزيع استجاباتها للمُدخلات المتنوعة.

ويفترض نموذج تحليل الطبقة الكامنة قدرة بنية الطبقة الكامنة تفسير أي ترابطات

بين الاستجابات في البيانات. ويعني ذلك، افتراض أن تكون الاستجابات لمدخلات متنوعة، داخل الطبقات الكامنة، مستقلة. وكما سبق لنا الإشارة إلى ذلك، على الباحث تحديد عدد الطبقات قبل التحليل. ولكن كيف يتسنى لنا معرفة قيامنا باختيار العدد «الصحيح»؟ عموماً، يجرب الباحثون أعداداً مختلفة من الطبقات، ويقدرون الأنسب للنموذج (على مستوى الاحتمالية اللوغاريثماتية (Log-Likelihood)، أو معيار أكايكي للمعلومة، أو معيار بايز للمعلومة، أو G^2، أو إحصاء تناسبي آخر).

ومع ذلك، إن تحديد عددٍ ما، قبلياً للطبقات الكامنة، وعدد أنماط استجابة في البيانات، لا يمّكن من تحديد نموذج تحليل الطبقة الكامنة بالكامل؛ مما يعني أن تقديرات المَعْلمات المتعددة سيعطي الاحتمالية القصوى نفسها، أو بعبارة أخرى - هناك حلول متعددة لمشكل تحليل الطبقة الكامنة الأنسب على نحو مماثل. وهذا يعني أيضاً، عدم استقرار تحليل الطبقة الكامنة في أغلب الأحيان، كما يمكنها بلوغ حلول مختلفة جداً، إذا ما أخذنا بعين الاعتبار القيم الأولى المختلفة. من أجل هذا، إنَّ عدد الطبقات الكامنة الممكن تحديدها - مع الأخذ بعين الاعتبار البيانات المدخلة - مقيدة. وفي العموم، يعد نموذج تحليل الطبقة الكامنة الأفضل من حيث القدرة على تحديد أعداد أصغر للطبقات الكامنة.

للفصل في إمكانية تحديد نموذج ما بالكامل، من الضروري تجربة قيم أولى متعددة، وفحص إمكانية تقارب النتائج من الحلّ نفسه. وعموماً، إن نموذج تحليل الطبقة الكامنة ذي التناسب الأفضل غير محدد بشكل تام، من أجل هذا يمكن تجسيد أحد الحلول لهذا الأمر في إنجاز العديد من تحليلات الطبقة الكامنة باستخدام البيانات نفسها، وإيجاد معدلات الحلول. واعتباراً من الآن - مع ذلك - يبقى هذا قضية تطرح إشكالية ذات تحليل طبقة كامنة.

ومن المهم أيضاً الإشارة إلى أن نموذج تحليل الطبقة الكامنة معرضة للتقارب على المستوى المحلي بدلاً من الحدود العليا العامة (Global Maxima). ويمكن حلّ هذه القضية من خلال محاولة التوسل بقيم أولى مختلفة، ومراقبة إحصائيات الاحتمالية اللوغاريثماتية؛ إنها قضية، يمكن تناولها، أكثر مما يمكن تناول قابلية التحديد.

وبما أن تحليل الطبقة الكامنة عمّر لبعض الوقت، فإن عدداً من رزم البرمجيات الإحصائية تضم روتينات تحليل الطبقة الكامنة. إن لدى نظام التحليل الإحصائي (SAS) برنامجاً يدعى معالج تحليل الطبقة الكامنة (PROC LCA)، يقوم بإنجازه بسهولة كبيرة. أما «الستاتا» (Stata)، فلا يملك تحليل طبقة كامنة مبنية داخلياً، بل يوجد برنامج مولّد من قبل المستخدم، يمكن - لسوء الحظ - تشغيله فقط بنسخ الطبعة الخاصة أو المعالجة المتعددة للستاتا (Stata's SE or MP)، وليس فاصل الثقة (IC). ومن الممكن أيضاً استخدام حزمة غلام (Gllamm)، المولّدة من قبل المستخدم لإنجاز تحليل الطبقة الكامنة. وإن غولدن الكامنة (Latent Golden) لحزمة البرمجيات الكامنة، هي مصممة بالخصوص لتحليل الطبقة الكامنة ونماذج متغير كامن أخرى، وهي سهلة الاستخدام (User-Friendly).

ولدى «R» عدد من الحزم التي تنجز تحليل الطبقات الكامنة، بما في ذلك تحليل المكوّن المستقل (Lca)، والنموذج الخطي العام (gllm). (ونبين هنا كيفية إنجاز تحليل طبقة كامنة في «R» مستخدمين حزمة تحليل الطبقة الكامنة المتعددة (poLCA) (Linzer and Lewis). وسنستخدم هذه لتحليل استجابة البيانات انطلاقاً من المسح الاجتماعي العام (http://www3.norc.org//Gss+website).

لقد قمنا بإعداد البيانات في وقت مبكر، بحيث انتقينا ست أسئلة، من خلالها تم استفسار المشاركين في الاستطلاع عن شعورهم حول نفقة الحكومة على مواد متنوعة: البيئة، والجيش، والرعاية الصحية، والمدن، والجريمة، والعلم. إذا كان الجواب بـ

1. على كُلّ مادة، فيعني ذلك أن المبحوث (Respondent)، يرى عدم إنفاق الحكومة ما فيه الكفاية؛ في حين إذا كان الجواب بـ
2. فيعني ذلك، المعدل العام الذي تنفقه الدولة. وأما إذا كان الجواب بـ
3. فيدل ذلك على أن الدولة تنفق كثيراً. وقد ضمّنا أيضاً مادة تمزج سؤالين:
   - منْ منَ المبحوثين الذين صوتوا في العام 2008؟ أو
   - إذا لم تصوتوا، فلصالح من كنتم ستصوتون؟

وأخذت الإجابات رمز 1 بالنسبة إلى أوباما، والرمز 2 بالنسبة إلى ماكين، والرمز 3 بالنسبة إلى رأي آخر. ونحمّل البيانات، ثم ننزّل ونفعّل حزمة تحليل الطبقة الكامنة المتعددة في R على النحو الآتي:

```
Library (foreign)
gssdata<-read.dta («gss_s12.dta»)
attach (gssdata)
install.packages («poLCA»)
library (poLCA)
```

وبعد ذلك، نحتاج إلى ربط المواد التي سنستخدمها لإنتاج الطبقات الكامنة، وحفظها في موضع يدعى xs2. وتراجع النموذج عن متغير المتراضي، لأسباب ستصبح أكثر وضوحاً أدناه في نقاشنا حول انحدار الطبقة الكامنة.

```
Xs2<-cbind(envir,urban,welfare,army,crime,science,vote08) ~1
```

وبمجرد القيام بهذا، يمكن للباحث تشغيل البرنامج باستخدام السطر الواحد للرمز (أو الشفرة (Code)) التالي:

```
lcal<-poLCA (xs2, gssdata, nclass = 2, maxiter = 1000, graphs = FALSE)
```

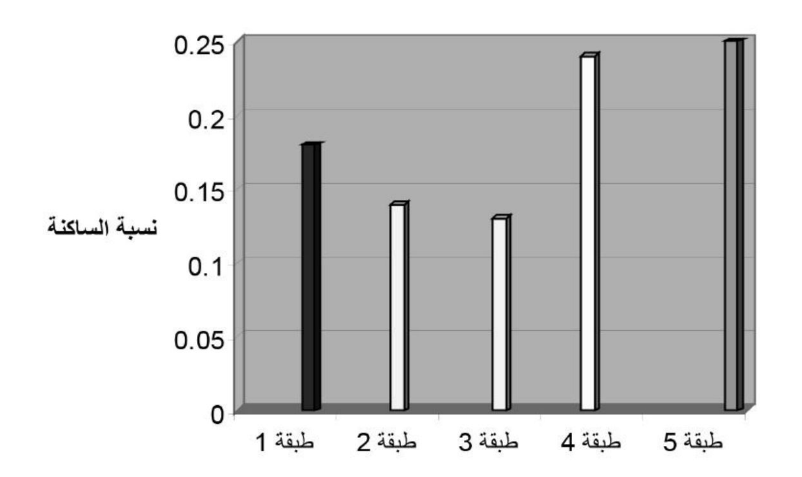
وكما ذكرنا آنفاً، إن xs2 هو الموضع الذي نحفظ فيه نموذج الطبقة الكامنة. وتُحدَّد البيانات بـgssdata ، ونحن بصدد تقدير نموذج ثنائي الطبقة (Nclass = 2). إن خيار الماكسيتر (Maxiter Option) يحدد العدد الأقصى للتكرارات لتعظيم الاحتمالية والرسوم البيانية = كاذب (FALSE) يقوم بإطفاء/ إيقاف الرسم البياني للدالة (Graphing Function). وستكون هناك الكثير من التكرارات لإيجاد حلّ الاحتمالية القصوى بالنسبة إلى نموذج ثنائي الطبقة. وعندما نزيد في عدد الطبقات، سنكون مضطرين للرفع من عدد التكرارات. ونشغل هذا النموذج عدداً من المرات بأعداد مختلفة من الطبقات الكامنة لإيجاد أفضل تناسبية للبيانات (الجدول رقم 1.13).

لاحظ عدم إشارة الإحصائيات التناسبية - بشكل لا يتطرق إليه الغموض- إلى عدد مثالي للطبقات. وإن معيار بايز للمعلومة يتجه من الخلف إلى الأعلى بعد خمس طبقات، في حين إن معيار أكايكي للمعلومة يأخذ مزيداً من الوقت بعض الشيء (بعد سبعة طبقات). أما بخصوص الإحصائية التناسبية التي تختار الاستماع إليها، فذلك شيء من قبيل دعوة الحكم في حضور أنصار الطرفين. ولغايات تتمثل في التقتير، نختار نموذج خماسي الطبقات.

يملك حلّ خماسي الطبقات انتشاراً طبقياً يشير إليه شريط الرسم البياني (Bar Graphs) في الشكل رقم 1.13. كما تصادف الطبقات 4 و5، و1 على نحو شائع جداً، بنسبة سكان تتراوح ما بين 20٪ و25٪. أما الطبقتان 2 و3، فهما أقل شيوعاً إلى حدّ ما.

الجدول رقم 13.1: انتقاء عدد الطبقات لأجل تحليل طبقة كامنة من خلال فحص الإحصائيات التناسبية.

رقم الطبقات	الاحتمالية اللوغاريثمية	معيار أكايكي للمعلومة	معيار بايز للمعلومة	G2
2	−13,945.89	27,953.77	28,126.92	2,483.06
3	−13,876.40	27,846.80	28,109.31	2,344.08
4	−13,821.94	27,769.88	28,121.75	2,235.16
5	−13,779.99	27,717.98	28,159.22	2,151.26
6	−13,751.00	27,692.01	28,222.61	2,093.29
7	−13,725.87	27,673.74	28,293.71	2,043.04
8	−13,714.06	27,682.12	28,391.45	2,019.45



الشكل رقم 1.13: توزيع حالات الطبقات الكامنة في تحليل الطبقة الكامنة.

الجدول رقم 2.13: احتماليات الاستجابة المشروطة بالطبقة انطلاقاً من تحليل الطبقة الكامنة.

		طبقة 1	طبقة 2	طبقة 3	طبقة 4	طبقة 5
البيئة	تزايد	0.2382	0.3196	0.6771	0.8043	0.7438
	اعتدال	0.3520	0.6307	0.2222	0.1830	0.2299
	تناقص	0.4098	0.0497	0.1007	0.0127	0.0263
المدن	تزايد	0.1536	0.1181	0.0000	0.4829	0.3666
	اعتدال	0.2786	0.6188	0.0429	0.4402	0.2815
	تناقص	0.4678	0.1415	0.6046	0.0362	0.1457
	غير متيقن	0.0999	0.1216	0.3525	0.0407	0.2062
الرعاية الاجتماعية	تزايد	0.2137	0.2264	0.7250	0.4738	0.5738
	اعتدال	0.2106	0.6541	0.1169	0.3097	0.2106
	تناقص	0.5757	0.1196	0.1580	0.2165	0.2156
الجيش	تزايد	0.4894	0.0866	0.1026	0.0438	0.4320
	اعتدال	0.3951	0.7151	0.3231	0.2609	0.5587
	تناقص	0.1155	0.1983	0.5744	0.6953	0.0093
الجريمة	تزايد	0.5378	0.3179	0.4800	0.4805	0.8441

	اعتدال	0.3512	0.6175	0.3176	0.4250	0.1387
	تناقص	0.1111	0.0646	0.2024	0.0945	0.0171
العلوم	تزايد	0.2945	0.2203	0.3384	0.4735	0.4033
	اعتدال	0.4606	0.7028	0.5091	0.4165	0.5347
	تناقص	0.2449	0.0770	0.1526	0.1100	0.0620
انتخابات 2008	أوباما	0.0643	0.6295	0.6204	0.8872	0.6015
	ماكين	0.8062	0.1952	0.1295	0.0646	0.2887
	آخر/ غير متيقن	0.1294	0.1753	0.2502	0.0482	0.1098

وتُعرَض احتمالات الاستجابة المشروطة في الجدول رقم 2.13. وكما تمت الإشارة إلى ذلك سابقاً، إن «معنى» الطبقات يحتاج إلى تأويل من لدن الباحث، ونسعى جاهدين للقيام بذلك هنا؛ فمعنى الطبقة 1 واضح جداً - وبتعبير بسيط، فهي تمثل المحافظين، الذين يفضلون الإنفاق على البيئة، والمدن، والرعاية الاجتماعية، والإنفاق أكثر على الجيش، ومحاربة الجريمة. وقد ساندوا ماكين (McCain) على حساب أوباما بهامش يصل إلى أكثر من 12 إلى 1. أما المجموعات الأخرى، فكلها من مساندي أوباما الذين بلغت نسبتهم - وهو الأمر الذي لا يدعو إلى كثير من الغرابة - حوالي 57٪ من مجموع الحالات. (لقد فاز أوباما بحوالي 54٪، وهذه المادة تضم مساندة الممتنعين عن التصويت). ولكن يحمل مساندو أوباما أولويات مختلفة. ويمكن وصف الطبقة 2 باعتبارها تضم وسطيين راضين (Satisfied Centrists)، إذ يرون إنفاق الحكومة معتدلاً في المجالات الستة كلها (ولو أنهم يرجحون أكثر، أفضلية الإنفاق على البيئة أو الجريمة أكثر من أشياء أخرى). أما أعضاء الطبقة 3، فيمثلون بشكل مثير للانتباه، الليبراليين المناهضين للمدينة (Anti-Urban Liberals). ويفضل هؤلاء الناس مزيداً من الإنفاق على البيئة، والفقر، والجريمة، ولكنهم يتجاوبون سلباً مع موضوع الإنفاق على المشاكل التي تواجه المدن؛ كما يفضلون الإنفاق على الجيش، ويساندون - أكثر من غيرهم في الطبقات الأخرى - الأطراف الثالثة، أو عدم اتّخاذ قرار. أما الطبقة 4، فتضم، التقدميين (Progressives)، بحيث يساندون أوباما على أعلى معدل، ويفضلون الزيادة في الإنفاق على قضايا بيئية. وهم وحدهم من يرون رفع معدل الإنفاق على العلوم، وعلى أي شيء آخر،

عدا الجيش. وأخيراً، يبدو أن الطبقة 5، تضم أناساً يفضلون، الإنفاق أكثر على مكافحة الجريمة، وتنظيف البيئة، والرعاية الاجتماعية. ويمكن اعتبار هؤلاء الليبراليين ممن يتبنى التوجه الحكومي، ذلك بأنهم يميلون - بشكل متزايد - إلى الإنفاق على الجيش أيضاً، في حين يعارض معظم مساندي أوباما هذا التوجه.

يعتبر هذا تحليلاً متسرعاً وقذراً للغاية. وبطبيعة الحال أنه بإمكاننا الحصول على طبقات متنوعة من خلال ضم متغيرات مختلفة؛ وتكون النتائج احتمالية وأكثر إيحائية منها قطعية. وداخل معظم الطبقات، كان يتخذ الناس مواقف مختلفة عن الموقف النمطي بشأن أي مادة معينة. علاوة على ذلك، تعد هذه النتائج غير مستقرة بخاصة؛ فعندما كنا ندير مزيداً من النماذج خماسية الطبقة ذات قيم أولى مختلفة، حصلنا على حلول مختلفة إلى حدّ ما. وفي كُلّ حلّ، هناك مجموعة محافظة واضحة، ذات احتماليات استجابة مماثلة جداً لتلك المذكورة أعلاه (على الرغم من أنها لم تكن تمثل دائماً الطبقة 1) ولكن تتنوع المجموعات التي تفضل أوباما من حيث ملفاتهم الشخصية المحددة.

انحدار الطبقة الكامنة

يعد انحدار الطبقة الكامنة امتداداً لتحليل الطبقة الكامنة، وهذا النوع من الانحدار لا يصنف فقط الحالات إلى عدد محتمل من الطبقات المحددة سلفاً، وإنما يستخدم أيضاً المتغيرات المشاركة (Covariates) لتنبؤ عضوية الطبقة. ويجعل منها هذا، مماثِلة للغاية لنموذج المعادلة الهيكلية، وتعمل بالأساس على النحو الذي يعمل به تحليل الطبقة الكامنة، مع انحدار متعدد الحدود (Multinomial) متصل به.

وفي R، من السهل جداً تحويل تحليل طبقة كامنة إلى انحدار طبقة كامنة، وذلك باستخدام بيانات المسح الاجتماعي العام (GSS) أعلاه. ولكن في هذه الحالة - ومن أجل التقتير - نفترض وجود فقط ثلاث طبقات كامنة.

```
Xs2<-cbind (envir,urban,welfare,army,crime,science,vote08)~AG
E+conserve+pared+inc
lcal<-poLCA(xs2,gssdata,nclass=3,maxiter=5000,graphs=FALSE)
```

لاحظ أن الاختلاف الوحيد في الرمز عن تحليل الطبقة الكامنة المنجز في وقت

سابق، هو تراجع المتغيرات الموقفية المحددة بالعمود بواسطة عن متغير اعتراضي. وإن برنامج «تحليل الطبقة الكامنة المتعددة» يؤول هذا الرمز باعتباره يحدد نموذج انحدار صفري (Null Regression Model). وعند إضافة المتغيرات المشاركة، تستخدم المتغيرات المحددة بالعمود في توليد احتمالية عضوية الطبقة، ثم تتراجع عضوية الطبقة عن متغيرات التنبؤ. إن انحدار الطبقة الكامنة يقدم لنا صورة، ليس فقط عن توزيع المواقف السياسية، وإنما أيضاً عن الترابطات الممكنة لكُلّ مجموعة، (الجدول رقم 3.13).

الجدول رقم 3.13: تقديرات المعلم المتنبئ لعضوية الطبقة في نموذج انحدار الطبقة الكامنة.

		Class 1	Class 2	Class 3
Environment	Increase	0.3130	**0.8401**	**0.6301**
	Just right	**0.3935**	0.1497	0.3521
	Decrease	0.2935	0.0101	0.0177
Cities	Increase	0.1677	**0.3645**	0.2757
	Just right	0.3058	0.2165	**0.5166**
	Decrease	**0.3972**	0.2071	0.1215
	Unsure	0.1293	0.2119	0.0861
Welfare	Increase	0.2771	**0.6846**	0.3689
	Just right	0.2619	0.1292	**0.4828**
	Decrease	**0.4609**	0.1862	0.1483
Military	Increase	0.4281	0.2521	0.0366
	Just right	**0.4394**	**0.3930**	0.4800
	Decrease	0.1325	0.3549	**0.4834**
Crime	Increase	**0.5621**	**0.7459**	0.3599
	Just right	0.3362	0.1726	**0.5606**
	Decrease	0.1017	0.0815	0.0795
Science	Increase	0.2914	0.4492	0.3710
	Just right	**0.5098**	**0.4619**	0.5464
	Decrease	0.1988	0.0888	0.0827
Election 2008	Obama	0.1547	**0.7401**	**0.8600**
	McCain	**0.6808**	0.1375	0.0574
	Other/unsure	0.1645	0.1224	0.0826

ويمنحنا الحلّ ثلاثي الطبقة، مجموعات تشكل 36٪، 31٪، و33٪ من السكان، على التوالي. وإن تشكيلة هذه المجموعات مختلفة إلى حدّ ما عن تشكيلة المجموعات التي نوقشت في حلّ تحليل الطبقة الكامنة خماسي الطبقة أعلاه. وتتألف الطبقة 1 من مزيد من الأفراد المحافظين، ممن يفضلون التراجع عن الإنفاق على الرعاية الاجتماعية و«المشاكل المدنية»، ويريدون في المقابل دعم مكافحة الجريمة. أما دعم الجيش قوي ولكن ليس قوياً مثل قوة الدعم الموجود في الطبقة 1 في الحلّ خماسي الطبقة. لقد دعموا ماكين بنسبة تصل إلى 68٪، ولكن ساند 15٪ منهم أوباما. وتضم الطبقتان 2 و3 ناخبين أكثر تقدمياً وأكثر وسطياً، على التوالي؛ في حين تفضل الطبقة 2 بشكل متزايد، الإنفاق على البيئة، والمدن، والرعاية الاجتماعية، وإيقاف الجريمة كما يعد ثلاث أرباع هذه المجموعة من أنصار أوباما. أما الطبقة 3، فهي فاترة بشأن الزيادة في الإنفاق، وتريد الإنفاق على الجيش. ولكن هذه المجموعة التي تبدو أكثر وسطية في آرائها – هي في الواقع أكثر دعماً على ما يبدو لأوباما.

نقدر عضوية الطبقة انطلاقاً من الدخل، والعمر، وتعليم الوالدين (أعوام)، وقياس المحافظة السياسية. ويتم قياس كُلّ المتنبئات بشكل مستمر. ينشأ قياس المحافظة السياسية من الاستجابات لسؤال يطلب الناس من خلاله ذكر أيديولوجياتهم السياسية، التي يتم ترميزها بسلم يتراوح ما بين 0 (ليبرالي جداً) و5 (محافظ جداً). ويجب أن يقرأ نتائج انحدار الطبقة الكامنة (الجدول رقم 4.13) بالطريقة نفسها التي تقرأ بها نتائج الانحدار اللوغاريثمي ذي الحدود المتعددة؛ أي إن انحدار الطبقة الكامنة من طبقة واحدة، مجموعة مرجعية، ويقدر العلاقة بين متغيرات المتنبئ والاحتمالات اللوغاريثمية في كُلّ طبقة من الطبقات الكامنة الأخرى بدل طبقة 1.

ومن الأهمية الإشارة إلى أنه على الرغم من أن لدى أولئك الموجودين في الطبقة 2 ملفاً شخصياً أكثر تقدمياً من أصل كُلّ المجموعات، فهم يشبهون إلى حدّ كبير الطبقة 1 (المحافظين) من حيث العمر، وتعليم الوالدين، والتوجه السياسي المبلغ عنه ذاتياً. إن الاختلاف الرئيس يتمثل في كون أولئك الموجودين في الطبقة 2 يتقاضون أجراً أقل من أولئك الموجودين في الطبقة 1. وفي المقابل، يختلف الأفراد

في الطبقة 3 اختلافاً كبيراً عن المحافظين في الطبقة 1 بطرق شتى - فهم أكثر شباباً، وإن لدى والديهم تعليماً محدوداً، ويبدو أنهم أفضل حالاً في المتوسط، ويحددون بكونهم أقل محافظة.

الجدول رقم 4.13: تقديرات المعلم المتنبئ لعضوية الطبقة في نموذج انحدار الطبقة الكامنة.

	طبقة 2 (مقابل 1)		طبقة 3 (مقابل 1)	
	معامل	P	معامل	P
العمر	0.002 (0.002)	.339	−0.008 (.002)	<.001
المحافظة	0.020 (0.001)	.151	−0.058 (0.014)	<.001
تعليم الوالدين	0.001 (0.001)	.902	−0.017 (0.036)	<.001
الدخل ($1000s)	−0.015 (0.001)	.001	0.011 (0.000)	.004
قار	0.002 (0.003)	.565	0.001 (0.000)	<0.001

ومهم أيضاً الإشارة إلى أن هذه المجموعة، تدعم أوباما بمعدلات مرتفعة بخاصة. وهذا يدل - على ما يبدو - على أن أولئك الموجودين في الطبقة 3 يشبهون شيئاً يسمى «حزب» الديمقراطيين، كما يعد العديد منهم أفراد راقين. وفي المقابل، تشبه الطبقة 2 شيئاً مثل التقدميين ذوي الياقات الزرقاء.

ويتطلب تأكيد هذه الأنماط تحليلاً أكثر كثافة مما يمكننا الانخراط فيه هنا. كان بوسعنا تناول الأسئلة الموقفية المختلفة بشكل عبثي، غير أن هذا التمرين يشير إلى كيفية استخدام انحدار التحليل الكامن عوض تجميع (Clustering) حضور البيانات المُدخَلة الثنائية أو الفئوية في الغالب.

نماذج مزيجة

ترتبط طبقة تقنيات تدعى نماذج المزيج (Mixture Models) إلى حدّ ما، بتحليل الطبقة الكامنة، وانحدار الطبقة الكامنة كليهما. وقد تم تطوير نماذج المزيج تصورياً في بداية الأمر، في العشرية الأولى من القرن الثامن عشر، إلا أنها لم تخضع للتجريب والممارسة بشكل كبير إلى غاية ظهور الحوسبة الحديثة. ولدى نماذج المزيج تطبيقات ضخمة في تحديد هوية المتكلم، وفي علم الوراثة، وفي تحليل

الصورة، وجدت أيضاً تطبيقاً في العلوم الاجتماعية، خاصة منذ تطور التقنيات، لتطبيقها على نماذج المزيج منذ تطور التقنيات، لتطبيق هذا النوع من النماذج على مسارات النمو (انظر مثلاً لوب (Laub)، وناجين (Nagin)، وسامبسون (Sampson)، .(1998)

وخلافاً لتحليل الطبقة الكامنة أو تحليل التجميع، يتم توجيه نماذج مزيج محددة نحو متغير نتيجة مهم جداً. وعموماً، توزع هذه النتيجة على نحو مستمر، سواء باعتبارها عادية، أو لوغاريثم عادي (Log–Normal) أو بواسون (Poisson)، أو غاما (Gamma)، أو ثنائية الحدود السلبية. والمفترض أن التوزيع المستمر الموجود في المتغير التابع، هو في الحقيقة، مزيج من توزيعين منفصلين عن سكان مختلفين. ومثال كلاسيكي على ذلك، هو الارتفاع بين عينة مكونة من رجال ونساء، حيث الجنوسة غير مرصودة. وإذا ما نظرنا إلى رسم بياني (Histogram) ما، فإن التوزيع سيكون إما عادياً أو ثنائي الحدين بعض الشيء؛ ولكن إذا أمكن لنا تحديد الجنسين بشكل منفصل، فسيكون بإمكاننا رؤية أن ما كنا نبحث عنه، هو - في واقع الأمر - توزيعان عاديان متداخلان، ومع ذلك، يتمثل مفتاح نموذج المزيج في عدم قدرتنا على رؤية - أو على أي حال، عدم قياس - التغاير الأساسي قيد البحث. ولكن لدينا ما يبرر - عادة نظرياً - اعتقادنا في أن العلاقة بين متغيرات المتنبئ والنتيجة تختلف عبر المجموعات الكامنة داخل ساكنة ما، إننا نتوقع رؤية تنوع المعاملات في نموذج انحدارنا، بشكل كبير بين الطبقات المختلفة. ويمكن لنماذج المزيج أيضاً نمذجة عضوية الطبقة، لتجعلها شبيهة جداً بانحدار الطبقة الكامنة.

وتوجد روتينات المزيج بالنسبة إلى العديد من النظم الإحصائية. وإن لدى نموذج الحزمة الإحصائية للعلوم الاجتماعية (SPSS) عقدة نماذج مزيجة خطية عامة. وقد كُتِبَت حزمات متنوعة في R، لأجل نماذج المزيج، بما في ذلك الفليكسمكس (Flexmix)، ولوغاريثمات نمذجة المزيج الغوسي (bgmm). ويشير «الغولد» الكامن إلى حزمة برمجيات متاحة تجارياً، ومخصصة تحديداً بالنسبة لنماذج متغير كامن، بما في ذلك نماذج المزيج. ولدى «الستاتا» (Stata) برنامج مستخدم مولّد، يدعى نماذج المزيج المحدود (fmm)، كما يمكن أيضاً استخدام برنامج لوغاريثمات نمذجة المزيج الغوسي ذي المستخدم المولّد.

ونبين نماذج المزيج مستخدمين برنامج نماذج المزيج المحدود في «الستاتا» (Deb 2012). وبما أن البرنامج برنامج مستخدم مولّد؛ أي إنه لا يشكل داخل الهندسة الأساسية لـ«الستاتا»، فإنه يحتاج إلى تحديد موقعه على شبكة الإنترنت أولاً.

findit fmm

وسيأخذك هذا إلى شاشة بحث عن برنامج قابل لتحديد موقعه بسهولة، ولهذا فما عليك إلا اتباع - ببساطة - التعاليم لتنزيله (Download). وتأخذ صيغة البرنامج الشكل الأساسي:

fmm depvar indvars [if] [in] [weight], components (integer) mixtureof (distribution) probability (model2) vce (type).

وفي هذه الصياغة، نخبر «الستاتا» بتقدير نموذج مزيج محدود، وحصر المتغير التابع في مجموعة من المتنبئات. ونحدد عدد المجموعات الكامنة التي نرى أنها مُمَثَّلة في البيانات (المكوّنات)، وخيار ميكستشر أوف (Mixture of) يسمح لنا بتحديد كيفية توزيع المتغير التابع (عادي، أو لوغاريثم عادي، أو «بواسون»، أو ثنائية الحدود السلبية، أو «أو غاما»). كما نستطيع أيضاً تحديد أشكال خطأ المعيار (vce)، مثلاً، قوي، «بوتسراب»، أو «الجاك نايف» (Jackknife). كما يسمح خيار الاحتمالية للمستخدم بتحديد المتنبئات لنمذجة احتمالية عضوية الطبقة.

ونحلل بياناتنا المستخلصة مرة أخرى، من المسح الاجتماعي العام، للعام 2012، مستخدمين كمتغيرنا التابع، مقياس مركب من التدين (Religiosity)، المكوّن من أجوبة عن أسئلة، يستفسر الأفراد فيها عن مدى أهمية ديانتهم إليهم وعن عدد المرات التي يصلون فيها، وعن عدد المرات التي يترددون فيها على الكنيسة. ثم نمزج هذه المواد في مقياس تصنيف محصل عليه، لديه كرونباخ (Chronbach) (ألفا) α لـ 0.81، مما يوحي - حقيقه - بارتباط المواد بشكل وثيق. ونقوم بنمذجة التدين، مستخدمين الدخل، والجنوسة، والعمر، وتعليم الوالدين، والعرق (الذي يرمز لها المسح الاجتماعي العام بأبيض، أو أسود، أو آخر؛ ونتخذ الأبيض، مجموعة مرجعية).

الجدول رقم 5.13: تقديرات المعلم بالنسبة إلى المتغير التابع (التدين)، باستخدام المربعات الصغرى العادية ونموذج المزيج (ثلاث مجموعات كامنة).

	المربعات الصغرى العادية		نموذج المزيج المحدود					
			فريق 1		فريق 2		فريق 3	
	معامل خطأ المعيار	p	معامل خطأ المعيار	p	معامل خطأ المعيار	p	معامل خطأ المعيار	p
الدخل	0.001 (.011)	0.997	-0.012 (.008)	0.146	-0.013 (.012)	0.228	0.005 (.005)	0.291
أنثى	0.282 (.037)	<0.001	0.035 (.022)	0.114	0.263 (.042)	<0.001	0.054 (.021)	0.011
عمر	0.010 (.001)	<0.001	0.0007 (.0006)	0.244	0.0060 (.0013)	<0.001	0.0001 (.0007)	0.867
تعليم الوالدين	-0.015 (.005)	0.003	-0.005 (.003)	0.115	-0.010 (.005)	0.055	-0.005 (.003)	0.084
أسود (مقابل أبيض)	0.471 (.054)	<0.001	-1.893 (.052)	<0.001	0.538 (.064)	<0.001	2.514 (.041)	<0.001
آخر (مقابل أبيض)	0.180 (.065)	0.006	0.060 (.048)	0.207	0.263 (.063)	<0.001	0.046 (.035)	0.178
قار	-0.547 (.143)	<0.001	1.184 (.104)	<0.001	-0.359 (.154)	0.020	-1.412 (.072)	<0.001

ويأخذ النموذج المحدد، الصيغة التالية:

xi : fmm religiosity lninc female AGE par_ed i.RACE, components (3) mix(normal) probability (EDUC AGE lninc female)

نحن بصدد تحديد نموذج تكون فيه النتيجة مزيجاً من ثلاث توزيعات عادية. وكما هو الحال بالنسبة إلى تحليل الطبقة الكامنة، إن عدد المكوّنات تختار - إجمالاً - إما بسبب معرفة أو نظرية قبْلية، أو لاختيار عدد المكوّنات التي لها أفضل تناسبية إحصائية. ونحن نختار الاستراتيجية الأخيرة.

ونمذج أيضاً عضوية المجموعة، مستخدمين، التعليم، والعمر، والدخل، والجنوسة. ومن أجل المقارنة، نبين أيضاً نتائج نموذج انحدار المربعات الصغرى العادية، الذي سيبين متوسط النتائج بالنسبة إلى الطبقات الكامنة الثلاث (الجدول رقم 5.13).

إن نموذج انحدار المربعات الصغرى العادية يخبرنا بأن التدين الأكبر لا علاقة له بالدخل، لكن هناك نسبة أعلى (في المتوسط) بين النساء مقارنة بالرجال، وأعلى بين السود وأفراد عرق آخر من البيض. وثمة علاقة سلبية بين تعليم الوالدين والتدين، مما يوحي بأن الوالدين المتعلمين بشكل أفضل، يميلون إلى تربية الأولاد تربية دينية أقل. وجدير بالاهتمام، تشغلينا لنماذج، ضمت التعليم، العديم الصلة بالتدين، سواء خضع تعلم الوالدين للرقابة أم لم يخضع. و**أخيراً**، ثمة علاقة إيجابية بين التدين والعمر.

ويصنف نموذج المزيج، الساكنة إلى ثلاث مجموعات أساسية مختلفة. ويبقى الدخل غير مرتبط بالتدين في كُلّ المجموعات الكامنة. أما العمر، فيعد متنبئ تدين إيجابي، غير أنه مهم إحصائياً فقط بالنسبة إلى المجموعة 2. وإن تعليم الوالدين مرتبط بقلة التدين، ولكن هذه النتائج مهمة فقط في $p < .10$ في مجموعتي 2 و 3.

وتعد النساء في مجموعتي 2 و3 أكثر تديناً في المتوسط، إلا أن الفرق لم يبلغ درجة الأهمية في المجموعة 1. أما الفوارق العرفية في التدين، فهي لافتة للنظر بشكل كبير؛ إذ يلاحظ في المجموعة 1، أن السود أقل تديناً بشكل كبير من البيض، وأن لا أهمية للفرق بين البيض والآخرين، ولكن في المجموعتين 2 و3، يعد السود أكثر تديناً في المتوسط، وهذا الفرق كبير بخاصة في المجموعة 3. أما أفراد من مجموعات عرقية «أخرى»، فهم شيئاً ما أكثر تديناً من البيض فقط في المجموعة 2.

وإن الجدول رقم 6.13، يوضح نتائج نماذجنا الاحتمالية لعضوية الطبقة. ومرة أخرى، لابد من أن يفسر هذا، على النحو نفسه الذي يفسر به انحدار لوجيستي متعدد الحدود. ويبدو أن هناك احتمال متزايد للتواجد في المجموعتين 2 و3 (عوض مجموعة 1) إذا كان المرء أكبر سناً، وهذه العلاقة أقوى بالنسبة إلى المجموعة 2 من المجموعة 3.

	Class 2 (vs. 1)		Class 3 (vs. 1)	
	Coeff. (SE)	*p*	*Coeff. (SE)*	*p*
Age	.037 (.007)	<.001	.018 (.006)	.002
Female	.412 (.216)	.056	.352 (.188)	.062
Education	–.073 (.038)	.051	–.111 (.033)	.001
Income	.102 (.076)	.177	.015 (.056)	.790
Constant	–1.659 (.894)	.063	2.153 (.680)	.002

ويرتبط التعليم سلبياً بالعضوية في أي من المجموعتين المرتبطتين بالمجموعة 1، إلا أن هذه العلاقة مهمة فقط في p <50.0 بالنسبة للمجموعة 3. وأخيراً، يبدو أن هناك علاقة إيجابية بين كون الفرد مؤنثاً، واحتمال تواجده في المجموعتين 2 و3، ولكن لهذا دلالة في p <.10.

خلاصة

لقد فحصنا في هذا القسم ثلاث تقنيات مَعْلمية (Parametric) لدراسة حضور المجموعات الكامنة في البيانات. وهذه الطرق - تحليل الطبقة الكامنة، وانحدار الطبقة الكامنة، ونمذجة المزيج - يمكن اعتبارها بدائل معلمية للتجميع. ويتوقف اختيار التقنية في القسم الأكبر على نوع بيانات المُدخل المتوافرة لدينا (مستمر أو فئوية)، وعلى مدى رغبتنا في تقدير عضوية المجموعة في صلتها بمتغير نتيجة معينة. وبينا إمكانية استخدام تحليل الطبقة الكامنة لفحص بيانات الاستجابة السياسية ضمن مجموعات تشترك في نمط التفكير، وإمكانية فحص نماذج المزيج للتغاير الأساسي في التدين. ومع ذلك، فإننا لا ننصح بالتأويل الذي يفيد بوجود هذه النماذج لطبقات أو مجموعات كامنة «حقيقية»؛ فهي بدلاً من ذلك، طرق، من خلالها يمكننا نمذجة النمط في البيانات إحصائياً، وهذا يمكن أن يكون مثمراً بالنسبة إلى تطوير النظرية والسؤال.

الفصل الرابع عشر

قواعد الارتباط

يعد التنقيب في قواعد الارتباط، إحدى أهم تقنيات التنقيب في البيانات المستخدمة بشكل واسع. واستخدمت في شكلها الكلاسيكي - وكما طورها في البداية، كُلّ من أغراوال (Agrawal)، وإيميلينسكي (Imieliński)، وسوامي (Swami) (1993) - في فحص بيانات سلة السوق في الخلفيات التجارية. وقد صمم هذا التطبيق العملي ليستفيد منه تجار التقسيط (Retailers)، المهتمين بأنماط ابتياع التي ينخرط فيها الزبائن. ولدى المحلات التجارية مجموعة معينة من المواد المعروضة للبيع في وقت محدد، بحيث يقتني الزبائن مجموعة من هذه المواد عندما يأتون إلى المتجر. وقد يرغب بائع التقسيط في معرفة مزيد من المشتريات التي يميل الزبائن إليها لدى شرائهم الحليب، أو البيض، أو بسكويت الكلب. ويمكن أن يساعد فهم هذه الأنماط، باعة التقسيط على بيع مزيد من البضائع، من خلال اقتراح - مثلاً - أن المواد التي تباع بكثرة، مخزّنة بالقرب من بعضها بعضاً. ويتمثل المشكل في كون أن محلات السوبر ماركت يمكن أن تتعامل مع عدد كبير من المعاملات التجارية، وتنقل منتوجات مختلفة كثيرة، بحيث يمكن بيع عدد هائل - إلى حدّ ما - من المواد في كُلّ معاملة تجارية.

ومن ثم، أضحى واضحاً أن هذا مشكلة بيانات ضخمة (A Big Data Problem)، بما أنه مشكلة يُمنع مجاله العام على المحللين من البشر. ومهما يكن، إذا باع محل تجاري ما 20 مادة منفصلة، ونحن مهتمين بالترابطات القائمة بين مادتين فقط، فإن

هناك 190 مزيجاً ممكناً. وإذا ما بحثنا في كُلّ الترابطات الممكنة (ليس في اتّجاهين فحسب) بين هذه المواد، فسيكون عدد الترابطات الممكنة، $2^{20} - 1 = 1.048.575$.
وبطبيعة الحال، لن تحدث معظم هذه المجموعة الممزوجة المحتملة من المبيعات. ومع ذلك، فإنه بين التجميعات نفسها التي تحدث بالفعل، توسم مشكلة البيانات بالحدة الشديدة للغاية إلى درجة استدعاء الآلية.

ومن أجل مناقشة الترابطات الأكثر أهمية، والترابطات التي يمكن تجاهلها، نحتاج أولاً إلى تقديم مصطلحين:

الأول يتعلق بالدعم (Support)؛ فدعم مزيج مادة معينة يعادل عدد كُلّ المعاملات التجارية التي تضم هذا المزيج، مقسوماً على مجموع عدد كُلّ التعاملات التجارية. إذن، إذا كان الدعم بالنسبة إلى مجموع المواد (حليب، بسكويت) هو 10٪ من أصل كُلّ المعاملات التجارية (معاملات قد تضم أي عدد من مواد أخرى).

وأما القياس الموالي، فهي الثقة (Confidence). وتشير ثقة قاعدة ما إلى احتمال رؤية مادة (Item) ما، مع الأخذ بعين الاعتبار رؤيتنا للمادة الأخرى. ومع ذلك، يقتضي هذا القياس منا اعتبار مجموعة فرعية من المواد في مجموعة موادنا، لاحقة (Consequent)، التي تعد نظيرة المتغير التابع، في حين نعتبر الآخر، المتبقي السابق (Antecedent).

وإن ثقة 75٪ في العلاقة (حليب ← بسكويت)، تعني أن في الوقت الذي يشتري فيه الزبون الحليب، فهو يشتري البسكويت أيضاً. ولاحظ - مع ذلك - أن قلب السهم يمكن أن يقدم ثقة مختلفة جداً؛ أي إن احتمال حصولنا على الحليب على اعتبار أن لدينا البسكويت يعادل على الأرجح، احتمال حصولنا على البسكويت على اعتبار أن لدينا الحليب. ومن المهم أيضاً الإشارة إلى إمكانية أن يكون كُلّ من اللواحق والسوابق مجموعات فرعية من مادة متعددة. ولهذا، من الممكن أن تكون لدينا قاعدة من قبيل (نقانق، كعك ← كاتشب، خردل) - احتمال حوزتنا على كُلّ من الكاتشب والخردل على اعتبار أن لدينا النقانق والكعك.

ويحيلنا هذا على شيء مهم لإنتاج قواعد الترابط المفيدة بالنسبة إلى البحث في

العلوم الاجتماعية. ومن الممكن تعيين مادة ما بصفتها هدفاً (Target)؛ أي تحديدها كلاحق. وستجد خوارزمية التعدين (Mining)، قواعد تشير إلى احتمال ذاك اللاحق، مع الأخذ بعين الاعتبار ظهور مواد سابقة.

وفي ضوء وجود عدد هائل من التجميعات (Combinations) في مجموعة البيانات، يجب تنفيذ قاعدة مميزة ما جديرة بالملاحظة، انطلاقاً من قواعد غير ذات صلة. ويتم القيام بهذا - نوعاً ما - بشكل عشوائي من قبل الباحث الذي يدير قاعدة خوارزمية «التعدين». ويختار الباحثون الحدّ الأدنى من قيم الدعم، أو الثقة، أو هما معاً، واستبعاد تجميعات بصفتها غير مهمة إذا ما فشلت في الاستجابة إلى الحدّ الأدنى من المعايير. وهناك طريقة أخرى للحد من عدد القواعد، المتمثلة في تحديد الحدّ الأقصى (أو الأدنى) لحجم مجموعات مادة اللاحق والسابق. **وأخيراً**، نشير إلى أن الفعل الحقيقي لتعيين مادة ما باعتبارها لاحقة، لها تأثير تقليص مجموعة القواعد العائدة.

من القضايا التي تثار في بيانات المعاملات التجارية، هو أن بعض المواد تباع في كثير من الأحيان (الحليب)، في حين تباع المواد الأخرى على نحو نادر (ملاعق). وأي حدّ أدنى لقواعد الدعم والثقة، سيضم - بالضرورة - قواعد كثيرة تحتوي على الحليب، وقواعد قليلة جداً تحتوي على الملاعق. ويمكن للمرء اعتبار ذلك نظيراً - في قواعد الترابط - لمشكل ذي نتائج نادرة. ومن الحلول المطروحة لهذه المشكلة، هو السماح للحد الأدنى من الدعم من أن يتنوع عبر المواد - أي استلزام حدّ أقصى من الدعم بالنسبة إلى مجموعة مواد تحتوي على الحليب، مثلاً، ودعم منخفض بالنسبة إلى تلك المجموعة من المواد التي تحتوي على الملاعق.

الآن، أمضينا معظم الوقت في الحديث عن محلات السوبر ماركت، ودكاكين البقالة، والحليب، والبسكويت - وقد يكون هذا لا محالة مفيداً جداً بالنسبة إلى أصحاب السوبر ماركت. ولكن بماذا يفيد هذا علماء الاجتماع وباحثين آخرين؟ ولماذا يستوجب على الباحثين الاهتمام بقواعد الارتباط؟ وكيف يمكن استخدام التنقيب في قاعدة الارتباط لدعم بحثنا؟

إننا نؤمن بأن بإمكانية أن تكون قواعد الارتباط أدوات استكشافية قوية عندما يكون لدينا بيانات ذات متغيرات مستقلة (سمات). ويمكن للمرء استخدام التنقيب في قاعدة الارتباط في شكلها غير الخاضع للرقابة (أي دون متغير الهدف) للحصول على فكرة بشأن كيفية اشتغال الأشياء بعضها مع بعض، غير أننا نعتقد في أنه من المفيد أكثر بالنسبة إلى العديد من الباحثين، الحصول على متغير نتيجة ذي دلالة. وبمجرد تحديد هذا، يستطيع المرء استخدام التنقيب في قاعدة الترابط للبحث عبر متغيرات مستقلة عديدة لاستكشاف المتغيرات التي تميل إلى الترابط مع النتيجة (Outcome). كما يمكن القيام بذلك بشكل أسرع وأكثر نجاعة من إنتاج مصفوفة الارتباط (Correlation Matrix). ولكن من الأهمية بمكان، الإشارة إلى إمكانية أن يجد التنقيب في قاعدة الترابط، تجميعات الشروط (Combinations of Conditions) المترابطة بالهدف. ويمكن لهذه التجميعات الإشارة إلى وجود تأثيرات مهمة ذات نوع تفاعلي(Ragin, 2008).

التنقيب في قاعدة الترابط في مُنَمْذِج الحزمة الإحصائية للعلوم الاجتماعية

لقد كان التنقيب في قاعدة الترابط موجوداً منذ أكثر من 20 عاماً - عند هذه النقطة - وتستعمل بشكل كبير في سياقات تجارية. ونتيجة لذلك، ظهر عدد من التطبيقات، القادرة على القيام به، غير أنه ليس مدمجاً في البرمجيات الأكثر استعمالاً بشكل مألوف من قبل الباحثين (SAS ,Stata ,PSS). ومع ذلك، هناك حزمة كبيرة ومعقدة بالنسبة إلى R تدعى اللا قواعد (Arules)، كما أن منمذج الحزمة الإحصائية للعلوم الاجتماعية قادرة أيضاً على إنجاز التنقيب في قاعدة الترابط. ونبين استخدام قواعد الترابط في منمذج الحزمة الإحصائية للعلوم الاجتماعية أدناه باستخدام البيانات المستخلصة من مسح المجتمع الأميركي، والتركيز على الأفراد الذين يفتقرون إلى تغطية التأمين الصحي.

من المهم إعداد بياناتك قبل التنقيب في القاعدة، ويفترض التنقيب في القاعدة - عادة - إن بياناتك موجودة في شكل المعاملة التجارية، حيث يمثل كُلّ سطر مزيجاً من مادة زبون ما أو بشكل أكثر دقة، مزيجاً لمادة معاملة تجارية. وإن المواد المتعددة التي تم شراؤها كلها، لا تظهر في السطر نفسه بل في السطور المتتابعة. وستكون

مجموعة البيانات (Dataset) طويلة وضيقة جداً، مع وجود عمود (Column) واحد، يدل على هوية المعاملة التجارية أو المشتريات، وعمود آخر يحدد منتوجاً فردياً. ومع ذلك، من الممكن إدخال بيانات في شكل جدول (Tabular Form)، وتمثل السطور هنا معاملات تجارية أو مشتريات، بحيث يشير كُلّ عمود إلى مادة يمكن شراؤها، ويضم عرض جدولي للبيانات (Tabular Data)، إذن، متغيرات وهمية، تعادل 1 إذا تم شراء المادة في معاملة تجارية معينة، وتعادل 0 إذا حدث العكس.

أما بالنسبة إلى علماء الاجتماع، فيعني حاجة التنقيب في قاعدة الترابط إلى متغيرات وهمية، أو على الأقل متغيرات فئوية، وعدم قدرتها على معالجة المتغيرات المستمرة بالنسبة إلى السوابق أو اللواحق. ولهذا، يجب أن تتحول المتغيرات المستمرة إلى متغيرات فئوية بواسطة طريقة من طرق التفريد/ التمييز (Discretization) قبل تشغيل روتين قاعدة الترابط. علاوة على ذلك، لا يقوم التنقيب في قاعدة الترابط بهذا على نحو جيد مع المتغيرات الفئوية المتعددة ذات الفئات العديدة، وسيكون - أصلاً - لدى هذه الفئات، معدلات دعم منخفضة تقريباً. ولهذا، على المرء اعتبار تجميع هذه الفئات ضمن فئات أوسع. ولمعرفة كيفية القيام بهذا، انظر إلى أقسامنا السابقة التي تناولت المتغيرات المستمرة المميزة، وتجميع المتغيرات الفئوية المتعددة.

وعموماً، نفضل الحصول على بيانات يظهر فيها فقط المتغيرات الوهمية، التي يمكن - مع ذلك - إنتاجها بطرق مهمة ومبتكرة. وتذكر، أنك لست بصدد بناء نموذج انحدار، ولهذا، لا يوجد داعٍ للتيقن من أن الفئات حصرية وشاملة بشكل متبادل، لعدم حاجتك إلى تأويل معاملات التأثير. ويجب اعتبار المتغيرات الوهمية، متغيرات مؤشر (Flag Variable) من أجل شروط مهمة. ولهذا، أمكن للمرء إدخال مؤشر ما من أجل مجموعة من الشروط - كون الفرد يتجاوز سنّ 30، ومسجل بصفته طالباً جامعياً في الكلية، مثلاً - من دون أن يقلق حيال طبيعة المجموعة المرجعية بالنسبة إلى هذا المؤشر.

أما في روتين ما من روتينات التنقيب في قاعدة الترابط، فهذا لا يطرح إشكالاً، إذ يمكن للمرء أيضاً إدخال متغيرات وهمية بالنسبة إلى تجميعات من الفئات

المتداخلة؛ فإن كان لدينا خمس مجموعات إثنية في بياناتها، مثلاً، أمكن للمرء إدخال متغيرات وهمية بالنسبة إلى كُلّ مجموعة على حدة (إدخالها برمتها، دون إقصاء المرجعية). كما يمكن إدخال كُلّ هذه التجميعات في وقت واحد في مجموعة بيانات التنقيب في قاعدة الترابط، وسيقوم الروتين - ببساطة - بخلط كُلّ متغيرات المؤشر المحددة، بحثاً عما يمكن اعتباره قواعد تنبؤية لمتغيرنا الهدف (المستمر).

إذن، قمنا بإنتاج مجموعة بيانات انطلاقاً من مسح المجتمع الأميركي، حيث ميزنا فيه العمر ودخل العائلة ضمن مجموعة فئات، وأنتجنا عدداً من المؤشرات (Flags) من أجل مجموعات من الحالات المهمة (مثل كون الفرد بالغ في سنّ العمل، وليس ضمن القوى العاملة). وأما شرطنا السابق، فهو «الافتقار إلى» حالات التغطية الصحية (Nohealthins)، وهو مؤشر يدل على الافتقار إلى أي تغطية صحية. نحن بصدد البحث عن صفات ومجموعة من الصفات المركبة، التي تعد سوابق (متنبئات) متكررة للافتقار إلى تأمين صحي. والآن، من الأهمية التأكيد على وجود قواعد تتنبأ بعدم الحصول على تأمين، عوض صفات الشروط التي - في الغالب - ما ترافق كون المرء غير مؤمن، وهذا فرق مهم لا محالة. وبتعابير رياضية، ستخبرنا قياسات الثقة التي سنجدها - مثلاً - عن احتمال افتقار المرء إلى التأمين الصحي، على اعتبار أنه فقير، وينحدر من أقلية عرقية، وليس عن احتمال كون المرء فقيراً وينحدر من أقلية عرقية، باعتبار عدم امتلاك أي أحد تأميناً صحياً.

ومن المهم الإشارة إلى الاحتمال الأساسي لمتغيرك الهدف أو النتيجة، عند تحديدنا للحدّ الأدنى من الدعم. وإذا حدث متغير هدفك بشكل نادر في بياناتك، فعلينا تحديد قاعدة الحدّ الأدنى من الدعم على نحو منخفض جداً، في الواقع. وستكون دوماً في حاجة إلى تحديد الحدّ الأدنى من الدعم على نحو أقل انخفاضاً من التكرار الأساسي لقيمة إيجابية في نتيجتك، وإلا لن تجد أي قواعد تذكر، كما يمكن تحديد قياسات الحدّ الأدنى من الثقة أكثر في تقديركم - حسب ما تراه مهماً كاحتمال شرطي.

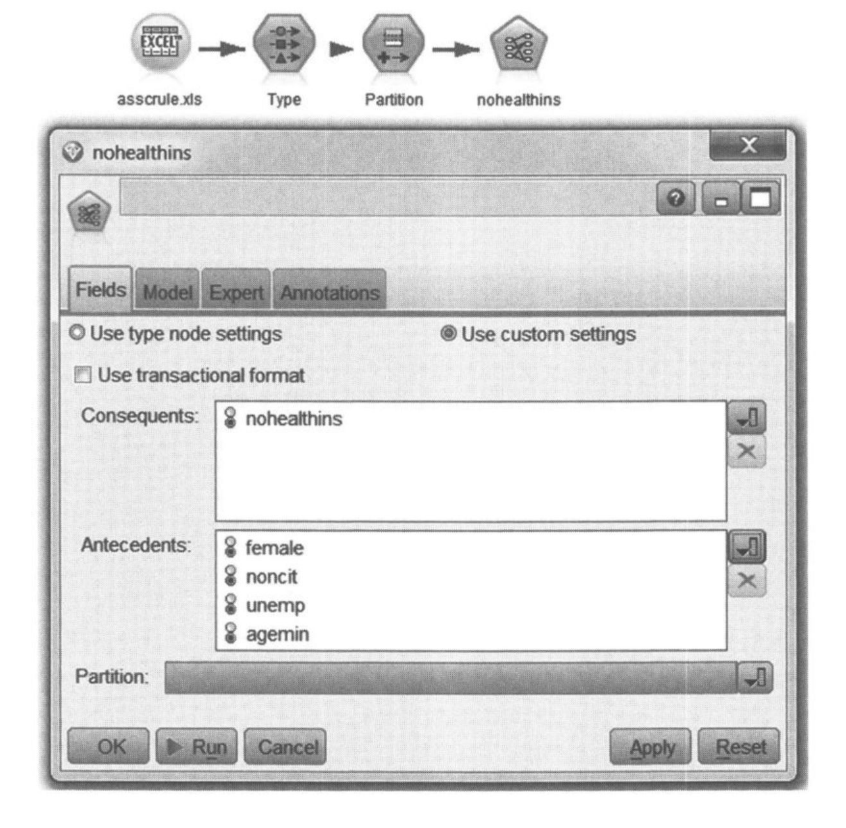
إن لقطة الشاشة أعلاه تشير إلى كيفية القيام بهذا التحليل. وإن العقدة الموجودة في أقصى اليسار، هي عقدة مصدرنا (Source)، التي قمنا فيها بانتقاء بياناتنا. وبعدها،

توجد عقدة «اشتق» (Derive) (لوحة المجال) (Field Palette)، حيث قمنا بتحويل بعض من متغيراتنا. وأخيراً، يتم انتقاء عقدة الفرضية[1] (apriori Node) من لوحة النمذجة (Modeling Palette) (الشكل رقم 1.14). وإن لدى المُنَمْذِج ثلاث روتينات قاعدة ترابط منفصلة. ونختار الفرضية (Apriori) لأنها تسمح للباحث بتحديد متغير مؤشرها (Flag). (ويولّد روتين الكارما (Carma) كُلّ القواعد الممكنة من دون إمكانية تحديد مؤشر ما، ويضع التسلسل (Sequence)، النظام حيث تكون المواد المدرجة فيه مهمة).

ثم نضع المَعْلمات، وفي هذه الحالة، يُنصح بالحفاظ على الحدّ الأدنى للدعم السابق منخفضاً نسبيا (في 0.5٪) لأن نتيجتنا تظهر فقط في حدود 14٪ من الوقت، ولكن نُبقي على الحدّ الأدنى من الثقة القاعدة عالٍ إلى حدّ ما. وفي إكسبيرت تاب (Expert Tab)، يمكننا إدخال إعدادات بديلة لفرز قواعدنا، كما يمكننا اختيار الإقصاء على أساس الاختلاف المطلق بين الثقة التي تمنح القاعدة والثقة القبْلية (مثلاً، احتمال رصد اللاحق (Consequent) بغض النظر عن السابق (Antecedent). ومن جهة أخرى، نستطيع القيام بعملية الاختيار استناداً إلى معدل هذين القياسين من قياسات الثقة. وفي المُنمذج (Modeler)، يُدعى هذان القياسان «اختلاف الثقة» (Confidence Difference)، ومعدل الثقة (Confidence Ratio)، على التوالي. ويعد هذان الإعدادان مناسبين بخاصة عندما تكون نتيجتنا نادرة نسبياً، كما هو الحال في هذه الحالة. وهناك طرق ممكنة أيضاً؛ «فاختلاف المعلومة» (Information Difference) يخبرنا بمدى تقديم ظهور السوابق لظهور اللاحق. ويأخذ بعين الاعتبار الدعم بحيث يتم تفضيل مزيداً من القواعد التي تحدث مراراً. ومع ذلك، فاختلاف المعلومة أقل فائدة من نتائج نادرة مثل نتائجنا. كما يقوم مربع كاي المعياري (Normalized Chi-Square) أيضاً على الدعم.

---

(1) تشير كلمة فرضية (apriori) إلى الخوارزمية الخاصة لتوليد قواعد الترابط، المستخدمة من قبل مُنَمْذِج الحزمة الإحصائية للعلوم الاجتماعية. وتعد خوارزمية الترابط الأولى التي تم اقتراحها من قبل أغراوال (Agrawal) وزملاء في المقال الأول حول قاعدة الترابط المذكورة أعلاه. ومنذ ذلك الحين، تم تطوير العديد من الخوارزميات الأخرى. وجدير بالاهتمام، أن جميعها يقود إلى المجموعة نفسها من قواعد الترابط لدى تطبيقها على البيانات نفسها، على الرغم من استعمالها منطقاً متفاوت بعض الشيء. (المراجع)



الشكل رقم 1.14: تدفق قاعدة الترابط والعقدة الفرضية (Apriori) في مُنَمْذِج الحزمة الإحصائية للعلوم الاجتماعية.

إن بياناتنا والإعدادات التي اخترناها، ولّدت 16 قاعدة ترابط منفصلة، المشار إليها في الجدول رقم 1.14. ولدى قراءتنا لهذه القائمة وتأويلها، يصير من المهم تذكر الأمر الذي تستطيع قواعد الترابط القيام به، والأمر الذي تعجز عنه. فالطريقة التي تقرأ بها قاعد رقم 1 - مثلاً - هو أن «80٪ من غير المواطنين (Non-Citizens)، التي تتراوح أعمارهم ما بين 30-39، يفتقرون أيضاً إلى التأمين الصحي». وتعد قواعد الترابط لا معلمية، ولا تشمل أي شيء من قبيل الضبط الإحصائي. إن قواعد الترابط - كما يقترح ذلك اسمها، لا تسمح لنا باستنتاج السببية - لا نعرف العوامل السابقة، هذا إذا كانت هناك عوامل أصلاً - في القاعدة رقم 1، التي تقود الناس إلى

عدم حصولهم على التأمين. كما لا تطلعنا على مجموعات المقارنة (مثل المواطنين الذين تتراوح أعمارهم ما بين 40–49، ويحسبون على البيض، ولديهم شهادات جامعية).

الجدول رقم 1.14: قواعد الترابط المولّدة بواسطة الخوارزمية الفرضية لِمُنَمْذِج الحزمة الإحصائية للعلوم الاجتماعية.

رقم القاعدة	السابق	الدعم ٪	الثقة ٪	الرفع
1	غير مواطن + العمر بين 30–39 + لاتيني + تعليم < تعليم ثانوي	0.6	80.0	5.87
2	شمال جنوب المنطقة الوسطى + غير مواطن + لاتيني + تعليم < تعليم ثانوي	0.5	78.0	5.72
3	غير مواطن + العمر بين 30–39 + التعليم < تعليم ثانوي	0.7	77.14	5.66
4	بطالة + العمر بين 19–29 + لم يسبق له الزواج أبداً + ذكر + مواطن عند الولادة	0.65	76.92	5.64
5	لاتيني + العمر بين 19–29 + تعليم < تعليم ثانوي + لم يسبق له الزواج أبداً	0.51	76.47	5.61
6	عاطل + العمر بين 19–29 + لم يسبق له الزواج أبداً + ذكر	0.71	76.06	5.58
7	غير مواطن + لاتيني + تعليم < تعليم ثانوي + لم يسبق له الزواج أبداً + ذكر	0.58	74.14	5.44
8	لاتيني + العمر بين 19–29 + تعليم < تعليم ثانوي	0.73	73.97	5.43
9	العمر بين 19–29 + تعليم < تعليم ثانوي + لم يسبق له الزواج أبداً + ذكر	0.78	73.08	5.36
10	العمر بين 19–29 + تعليم < تعليم ثانوي + ذكر	0.95	72.63	5.33
11	عاطل + العمر بين 19–29 + ذكر + مواطن عند الولادة	0.73	72.60	5.33
12	عاطل + العمر بين 19–29 + ذكر	0.82	71.95	5.28
13	العمر بين 19–29 + لاتيني + تعليم < تعليم ثانوي	0.74	71.62	5.25

14	شمال جنوب المنطقة الوسطى + غير مواطن + تعليم < تعليم ثانوي	0.56	71.43	5.24
15	لا يوجد في القوة العاملة + غير مواطن + لاتيني + تعليم < تعليم ثانوي	0.52	71.15	5.22
16	العمر بين 19–29 + تعليم < تعليم ثانوي + لم يسبق له الزواج أبداً + ذكر + مواطن عند الولادة	0.61	70.49	5.17

ولكن قواعد الترابط مفيدة جداً في إخبارنا بمن يفتقر إلى تأمين صحي. علاوة على ذلك، فهي تقوم بذلك على نحو متعدد المتغيرات مثير للاهتمام. ويسمح لنا المُدخل بالقيام ببعض المقارنات المحلية. دعنا نقارن قاعدتي 1 و3؛ فالفرق الوحيد هنا يتمثل في كون قاعدة رقم 1 أكثر دقة إلى حدّ ما، بما أنها تضم «واللاتينيين» (and latino). أما المؤشرات الأخرى، فهي متطابقة، ولو أن ثقة القاعدة بالنسبة إلى قاعدة رقم 1 أعلى من قاعدة رقم 3. ويبدو أن من بين أولئك الذين تتراوح أعمارهم ما بين 30–39، وغير مواطنين، ولديهم شهادة تعليمية أقل من الشهادة الثانوية، يلاحظ أن اللاتينيين أقل إلى حدّ ما، من المتوسط للحصول على تأمين صحي. ولمعرفة ما إن كان هذا الفرق الأخير «له دلالة»، على المرء إنجاز اختبار إحصائي رسمي منفصل. وإن قيمة القواعد الترابطية، في هذه الحالة، هو أنه يمكن أن نقترح علينا نوع الاختبارات الرسمية – من بين مجموعات فرعية – التي قد تكون مهمة.

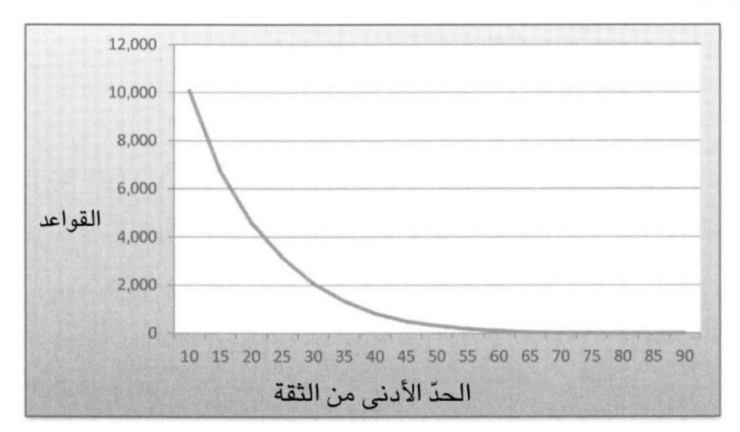
ويضم الجدول قياسات إحصائية قليلة. ويوجد في العمود الثالث الدعم (Support). وكل هذه المجموعات الفرعية تضاهي قسماً صغيراً من مجموع السكان، أي أقل من 1٪ من كُلّ الحالات. إن الافتقار إلى التأمين الصحي هو «حدث نادر» (ولو أنه ليس نادراً كما يجب)، وإن اتّخاذ مجموعة «الافتقار إلى تأمين صحي»، وعدد من المجموعات الأخرى صغير جداً في الواقع، بطبيعة الحال. أما العمود الموالي فيمثل الثقة، التي هي الاحتمالية الشرطية لعدم منح تأمين ما، مع الأخذ بعين الاعتبار الشروط السابقة (Antecedent Conditions). و**أخيراً**، لدينا الرفع (Lift)، الذي يشير إلى تحسين تنبؤ النتيجة التي نحصل عليها من خلال معرفة السوابق – أي الاحتمال البَعدي (Posterior) مقسوم على «الاحتمال القبلي». وفي الإنجليزية، تعد هذه – ببساطة – احتمال عدم الحصول على تأمين، مع الأخذ بعين الاعتبار مجموعة

من السوابق (مثلاً اللاتيني ممن يتراوح عمره بين 30-39)، وغير مواطن، وحاصل على تعليم أقل من التعليم الثانوي)، مقسوم على الاحتمال البسيط الذي يفيد عدم الحصول على تأمين في مجموع الساكنة. وإن معدل رفع 5.87، يخبرنا بأن المجموعة الفرعية ذات المصلحة، تصل إلى 5.87 مرة أكثر من احتمال الافتقار إلى تأمين صحي من متوسط الساكنة. ولأننا وضعنا الحدّ الأدنى للثقة في حدود 70، وأن معدلنا الأساسي للافتقار إلى التأمين هو 13.63٪، فإننا نرى فقط المجموعات الفرعية ذات رافعات تصل إلى 5.13 أو أعلى من ذلك.

نرى في هذا الجدول ظهور المؤشرات مراراً وتكراراً، ومردّ ذلك، أحياناً - إلى كون المؤشر مألوفاً ببساطة (مثل شراء الحليب أو الخبز، في بيانات محل البقالة). ولكن ليس هذا ما يحدث بشكل واضح في أغلب الأحيان هنا، ذلك بإن فئات الأغلبية، أو حتى الفئات المشروطة - ليست سائدة، وفي حالة من هذا القبيل، تعد النتائج أكثر أهمية وإفادة. كما يوضح الجدول بجلاء إخفاق النظام الحالي للتأمين الصحي - على الأرجح - في توفير ضمان صحي لمزيد من السكان المهمشين - أي من غير المواطنين بخاصة، وأولئك الحاصلين على تعليم رسمي قليل، والشباب، والعاطلين. وحيثما تصادفت مجموعة قليلة من هذه المؤشرات، ارتفعت معدلات الافتقار إلى التغطية الصحية بشكل لافت للنظر.

لاحظ أننا حددنا الحدّ الأدنى لثقة القاعدة في معدلات مرتفعة جداً، ويمكننا تحديدها في معدلات أقل بكثير، ولكن قد نحصل بعدها على مزيد من قواعد الترابط. وفي هذه البيانات، إذا حددنا الحدّ الأدنى للثقة في حدود 90، فلن نجد أي قواعد بالمرة. علينا أن نخفضه إلى 80 قبل إيجاد قاعدة مستقلة. ولكن بعد هذا، سيزيد مزيد من الانخفاضات في الثقة من عدد القواعد الموجودة، أضعافاً مضاعفة. وفي حدود 70، نجد (كما في المعطى أعلاه) 16 قاعدة. وفي حدود 40، يرتفع إلى 815، وهكذا. وأخيراً، إذا حُدّد الحدّ الأدنى للثقة في 15، فقط فوق متوسط الساكنة التي تفتقر إلى التأمين الصحي، فسنجد 6,729 قاعدة. ومن ثم، فإن تخفيض الثقة يُلقي بشبكة أكبر، ليسمح ذلك بجمع مزيد من المعلومات المفيدة حول مجموعات فرعية مهمة احتمالاً، ولكن على حساب إثقالنا بالمعلومات (الشكل رقم 2.14). وفي هذه

النقطة، تكون غاية تمرين التنقيب في البيانات قد انهزمت، وعوض استخدام القوة الحاسوبية لإبراز الأنماط المفيدة داخل قدر ضخم من البيانات، كان علينا تنظيم هذا القدر الهائل من البيانات على نحو مختلف، والعمل على بذل قليل من الجهد في سبيل التبسيط.



الشكل رقم 2.14: عدد من قواعد الترابط العامة عند مستويات مختلفة من الحد الأدنى للثقة.

إن التنقيب في قاعدة الترابط هو تقنية من تقنيات التنقيب في البيانات المستعملة غالباً في إعدادات تجارية، ولكننا بيّنا آنفاً إمكانية استعمالها بشكل مفيد من قبل الباحثين من أجل غايات استكشافية. إن قواعد الترابط يمكن أن تساعد الباحثين على استكشاف حالات وتجميعات من الحالات التي تحدث بشكل متكرر بالإضافة إلى نتيجة هدف معينة. وبينما لا تمنحنا القواعد المولّدة أي معلومة حول الطبيعة السببية للعلاقة بين السوابق واللواحق، فستبقى إيحائية جداً، ولكن بالإمكان أن تكون مثمرة في اقتراح مسارات مهمة في البحث، إذا ما استخدمت بالاشتراك مع طرق استكشافية أخرى.

استنتاج

ما هو القادم؟

لقد مر أكثر من نصف قرن على بداية انتشار الحوسبة عبر المجتمع، ليصبح تأثيرها بديهي في العديد من مناحي حياتنا. ولما شرعت الأعمال التجارية في تركيب الحواسيب في الستينيات والسبعينيات، تم ذلك وفق أهداف محدودة وعملية جداً في الذهن: العمل على «جعل» أنواع معينة متنوعة من سجلات المعاملات التجارية «آلية» من أجل تقليص نفقات إعداد الفواتير والاستعانة بالحسابات والميزانيات العمومية. ويدرك القليل أن إحدى المنتوجات الثانوية (By-Product) المهمة، قد تمثل طوفاناً من بيانات الأعمال التجارية التي تمكّن المديرين من الولوج إلى تفاصيل المبيعات أو تدفق المال في تلك اللحظة بالذات، عوض انتظار إغلاق الحسابات في آخر الشهر أو العام. كما أدركت الشركات سريعاً إمكانية تحليل آلاف التفاصيل من المعاملات التجارية لتحديد أجزاء الشركة ذات الأداء العالي والمنخفض، بغية تقليص حجم المخزون للتحول إلى توفير منتوج في الوقت المناسب، أو إلى مبيعات وإعلان أكثر دقة. إن بيانات المعاملات التجارية تغيرت من كونها عبء عمل ورقي إلى كونها مصدراً قيماً للمعلومة، واستبصاراً حول عمليات وعروض تجارية. إن عصر البيانات الضخمة قد بدأ.

لقد أصبحت مصادر معلومات وإمكانات جديدة متاحة للتحليل بما أن قدراً

كبيراً من الاتّصال تحول إلى اتّصال رقمي، أو انتقل عبر الإنترنت. نستطيع متابعة الأوبئة من خلال الاستفسار عن الأعراض عبر الإنترنت؛ واستكشاف مشاعر الرأي العام من خلال تحليل عدد الكلمات المدرجة في وسائل الإعلام؛ كما يمكن تعقّب التحولات في استخدام اللغة عبر وثائق غوغل، وفحص الشبكات الاجتماعية، وانتشار الأفكار؛ إضافة إلى (إذا كنا نمثل وكالة الأمن القومي) التجسس على المكالمات الهاتفية الرقمية، والبحث عن الإرهابيين أو عن «إبر في أكوام قش» أخرى من خلال البحث عن أنماط في مقدار ضخم من البيانات.

في بداية الأمر تم تطوير طرق التعليم الآلي، وطرق التعرّف على الأنماط على يد علماء الحاسوب، وعلماء الرياضيات التطبيقية لغايات عملية مثل التعرف على خطّ اليد، وعملية فرز آلية للبريد، والترجمة الآلية، والرؤية الروبوتية. ولكن امتدت هذه الطرق بسرعة إلى الطرق التي نحلل بها البيانات الكمية بشتى أنواعها. ونتيجة لذلك، أصبح التحليل والتنقيب في البيانات مجالين مزدهرين؛ فالتنقيب في البيانات مشروع توسّعَ بشكل سريع، ليعطي ميلاد تخصص جديد، يدعى «علم البيانات»، وتخصصات مهنية جديدة.

لقد كانت غاية هذا الكتاب تقديم مدخل ميسر إلى بعض من هذه الطرق. وبالنظر إلى تكاثر الدورات الدراسية حول التنقيب في البيانات وتحليلات المعاملات التجارية، نتوقع قرار العديد من الناس تعلم هذه الطرق الجديدة من أجل تحليل البيانات. ومن ذلك، نقر بأن هذا المجال من البحث لا يزال في مراحله الأولى، وتوجد أصلاً بعض الحواجز التي تعرقل تطوره في المستقبل. وليس مصدر هذه العراقل، مجتمعات علوم الحاسوب أو الرياضيات التطبيقية التي تعد مبتكرة بوضوح بنسبة استثنائية، لتنتج طرقاً وخوارزميات جديدة. ولكن، تبقى البرمجيات تطرح مشاكل. وغالباً ما يكتب مختصو التنقيب في البيانات الأذكياء برامجهم في المتالاب (MATLAB) أو البيثون (Python)، ولكن سيكون معظم المختصين في التنقيب في البيانات الطموحين، غير راغبين أو قادرين على إنتاج برنامج بدءاً من الصفر. وكما سيلاحظ القراء، إننا لجأنا - في المقابل - في هذا الكتاب إلى لوحة منتوجات سهلة الاستخدام بشكل معقول، ومتاحة على شكل واسع، لتقديم نظرة عامة عن طرق التنقيب في البيانات، أحياناً باستخدام «الغامب برو»، وأحياناً باستخدام منمذِج

الحزمة الإحصائية للعلوم الاجتماعية (SPSS)، وأحياناً أخر باستخدام R، وهكذا. وإن هذا التشظي لأدوات برمجية سهلة المنال - لعدم وجود حالياً أي حزمة مستقلة تغطي كُلّ الأدوات التي يحتاجها المرء - تنتج عبئاً لمختصين محتملين في التنقيب في البيانات. وهناك منحنى تعلم حاد (Steep Learning Curve) في الاستئناس بأنواع مختلفة جداً من البرمجيات.

إننا في بعض الأحيان بعيدين كُلّ البعد عن الإعجاب بجودة هذه المنتوجات، على الرغم من استخدام جميعها. وفي كثير من الأحيان، يتوقف البرنامج كلية عن الاشتغال، أو يشتغل دائماً. وقد تحدث هذه المشاكل عندما تكون مجموعات البيانات ضخمة: أكثر من ألف حالة. ويبدو من العبث كتابة كتاب، والمرء مفتوناً بالبيانات الضخمة، وبعدها تقديم أمثلة متوسلين بحالات لا تتجاوز المائة، وهو ما يجد المرء - مع ذلك - في العديد من الكتب في هذا الموضوع. لقد حاولنا اجتناب القيام بذلك، واستخدمنا في هذا الكتاب بيانات ذات حجم معتبر متى كان ذلك ممكناً، ولكن يجب على القراء أن يدركوا إمكانية أن يصادفوا إحباطات مماثلة لدى تطبيقهم التنقيب في البيانات على مجموعات بياناتهم الواسعة. نتمنى أن تخف حدة هذه المشاكل سريعاً كلما أصبحت منتوجات البرمجيات أكثر شمولية في الأدوات التي يضمونها، وكلما تيقن معدو البرمجيات من إمكانية معالجة مجموعات بيانات ضخمة. ولكن، في الآونة الراهنة لا تزال هذه المشاكل تشكل خطراً.

مازال تحليل بياناتنا، يتطلب في تجربتنا استبصارات وخبرة جمة لدى المحلل، على الرغم من تقديم التنقيب في البيانات أدوات آلية. ولا يمكن للمرء إدخال - ببساطة - بيانات أولية (Raw Data) داخل هذه البرامج، ويتوقع الحصول على أي شيء مفيد. إن خبرة المحلل حاسمة في تحديد المشكل أو السؤال المعالَج. وإن معالجة البيانات قبلياً - من خلال البت في المتغيرات التي نضم، وفي كيفية قياسها - هي مرحلة تستهلك الوقت والتفكير معاً. إنَّ التحاليل الاستكشافية للبيانات - من خلال البحث عن السمات والمتغيرات المهمة، والوقوع في الحيرة بسبب نتائج غير متوقعة أو الافتقار إليها أصلاً - تطرح إشكالاً دقيقاً؛ ذلك بأنه في العديد من الحالات تكون مسألة اختيار التقنية معقدة. هناك العديد من البدائل، وربما يريد المرء بدائل متعددة. وفي تجربتنا، تتحسن النماذج بشكل كبير بفضل الضبط (Fine-Tuning)

عبر التجربة والخطأ (Trial and Error)، وتعديل المَعْلمات. و**أخيراً**، تعد ترجمة النتائج من التحاليل إلى شيء يمْكن للعملاء التجاريين أو الزبائن فهمه، تعهداً غير تافه.

وبالتالي، كي يصبح المرء مختصاً في التنقيب في البيانات، تشمل الخطوة التالية - بعيداً عن إتقان مضمون هذا الكتاب - تطوير هذه الاستراتيجيات والمهارات عبر الانخراط الواسع في البيانات والمشاريع.

الثبت التعريفي

أشجار الانحدار والتصنيف (Classification and Regression Trees CART): هي طرق تعليم آلي من أجل تشكيل نماذج تنبؤ انطلاقاً من بيانات. ويتم الحصول على هذه البيانات بواسطة التقسيم العودي لحيز البيانات، والعمل على مواءمة نموذج تنبؤ بسيط داخل كُلّ تقسيم. ونتيجة لذلك، يمكن للتقسيم تمثيل شجرة القرار بيانياً.

أشجار القرار (Decision Trees): تشير إلى شكل بسيط وقوي من أشكال التحليل المتعدد المتغيرات، ويتم إنتاجها من قبل الخوارزميات التي تحدد طرقاً متنوعة من تقسيم مجموعة بيانات إلى قطع شبيهة بالفروع.

انحدار تدريجي (Stepwise Regression): إنه أداة آلية، تستخدم في المراحل الاستكشافية لبناء نموذج ما بغية تحديد مجموعة فرعية مفيدة للمتنبئات. ويضيف هذا الإجراء المتغير الأكثر دلالة أو يزيل المتغير الأقل دلالة خلال كُلّ خطوة.

انحدار الطبقة الكامنة (Latent Class Regression): يشمل انحدار الطبقة الكامنة تشكيل طبقات كامنة لمجموعات فرعية أو قطع غير مرصودة لحالات ما؛ أي إنه يربط مجموعة من المتغيرات المتعددة التباينات المرصودة بمجموعة متغيرات كامنة. إنه نوع من نموذج متغير كامن. ويدعى نموذج طبقة كامنة، لأن المتغير الكامن منفصل.

انحدار لوجيستي (Logistic Regression): هو أداة إحصائية، تروم التحليل المناسب للانحدار عندما يكون المتغير التابع ثنائياً، وهو تحليل تنبؤي مثله في ذلك مثل كُلّ أنواع تحليلات الانحدار. ويستعمل في وصف البيانات وتفسير العلاقة بين متغير تابع ثنائي ومتغيرات مستقلة عادية للغاية مثلاً. وهو صعب التفسير أحياناً.

انحدار متعدد (Multiple Regression): هو وسيلة إحصائية يهدف إلى التعرف أكثر على العلاقة القائمة بين متغيرات مستقلة أو متغيرات متنبئة عديدة وبين متغير تابع أو متغير معياري. وبمجرد تحديد هذه العلاقة، يكون بإمكانك الحصول على معلومات حول جميع المتغيرات المستقلة، واستخدامها في تشكيل تنبؤات أكثر قوة ودقة حول السبب الذي جعل من هذه الأشياء أن تكون على الشكل الذي هي عليه.

انحدار المربعات الصغرى العادية (Ordinary Least Squares Regression): هي طريقة إحصائية لتقدير المَعْلمات غير المعروفة في نموذج الانحدار الخطي بغية تقليص مجموع مربعات الاختلاف بين الاستجابات المرصودة، وهي قيم المتغيرات المتنبأة في مجموعة بيانات معينة، وبين تلك القيم المتنبأة من قبل دالة خطية لمجموعة متغيرات تفسيرية.

k-أقرب الجيران (k-Nearest Neigbours): هي إحدى خوارزميات التصنيف الأساسية في التعلم الآلي. وتستخدم في التصنيف ومشاكل انحدار تنبؤية.

تجريف البيانات (Data Dredging): وتدعى أيضاً «اصطياد البيانات»، وهي ممارسة التنقيب في البيانات حيث تحليل أحجام هائلة من البيانات للبحث عن علاقات ممكنة بين البيانات. وأما الطريقة العلمية التقليدية، فتبدأ بفرضية ما، وتُتبع بفحص للبيانات، على عكس تجريف البيانات التي تسعى إلى استكشاف أنماط أو ارتباط متغيرات، يمكن تمثيلها باعتبار أن لها دلالة من حيث الحصيلة الإحصائية، دون اقتراح فرضية محددة حول السببية الأساسية.

تجميعات الحاسوب (Computer Clusters): هي تجميعات تحتوي على مجموعة من الحواسيب المترابطة ارتباطاً وثيقاً، وتعمل معاً، فيتم اعتبارها نظاماً

واحداً. ولدى كُلّ عقدة من عقد تجميعات الحاسوب، المَهَمّة نفسها الموكلة إليها، بمراقبة من برمجية ما.

تحليل المكوّن الرئيسي (Principal Component Analysis) (PCA): هو تقنية تستخدم للتركيز على التباين، والوقوف عند أنماط قوية في مجموعة بيانات، وغالباً ما يستخدم أيضاً لتسهيل عملية استكشاف البيانات بشكل واضح.

تحليل المكوّن المستقل (Independent Component Analysis) (ICA): هو تقنية إحصائية وحاسوبية لاستكشاف عوامل خفية مؤسسة لمجموعات متغيرات عشوائية، وقياسات، وإشارات.

تصحيح بونفيروني (Bonferroni Correction): إنه تصحيح يعتمد المقارنة المتعددة المستخدمة عندما تنجز عدة اختبارات إحصائية تابعة أو مستقلة في آن واحد.

تعيلم آلي (Machine Learning): هو حقل فرعي من حقول علوم الحاسوب، التي تمنح الحواسيب القدرة على التعلم دون أن تكون مبرمجة بشكل واضح.

تقسيم الأشجار (Trees Partition): هي أداة التنقيب في البيانات النموذجية؛ فهي بسيطة، وفعالة، وتعتمد على البيانات، بشكل مطلق؛ إنها أولاً وقبل كُلّ شيء، مصنف، تستعمل خصائص المدخل لخلق نموذج يقسم حالات إلى فئات ذات قيم مختلفة على مستوى نتيجة ذات دلالة.

تقسيم عَودي (Recursive Partitioning): إنها خوارزمية، تشير إلى فكرة بسيطة جداً من التجميع، وهي عكس التجميع التراتبي، كما تشير إلى عملية متدرجة، تتشكل خلالها شجرة قرار ما بواسطة تقسيم أو دون تقسيم كُلّ عقدة على حدة إلى عقدتين شقيقتين.

حزمة إحصائية للعلوم الاجتماعية (Statistical Package for the Social Sciences) (SPSS): هي برمجيات تستخدم في التحليل الإحصائي لإدخال البيانات، وتمثيلها في بيانات وجداول. وهي قادرة على معالجة بيانات ضخمة.

التنقيب في البيانات (Data Mining) (DM): يطلق على مجموعة من تقنيات الحاسوب المكثف، بغية استكشاف البنية، وتحليل الأنماط في البيانات.

حيلة أو خدعة النواة (Kernel Trick): هي تقنية من تقنيات التعلم الآلي لتجنب حساب مكثف ما في بعض الخوارزميات التي تجعل الحساب يسير من كونه إجرائياً إلى كونه غير إجرائي.

مربع كاي للكشف عن التفاعل التلقائي (CHi-Squared Automatic Interaction Detector) (CHAID): يشير إلى خوارزمية، تستعمل من أجل استكشاف العلائق القائمة بين متغير الاستجابة الفئوية، ومتغيرات متنبئ فئوية أخرى. ويستخدم مربع كاي للكشف عن التفاعل التلقائي عندما نبحث عن أنماط في مجموعات البيانات ذات تغيرات فئوية كثيرة، وهو طريقة مناسبة لتلخيص البيانات باعتبارها علائق، يمكن رؤيتها بسهولة.

«رابدماينر» أو منقب سريع (RapidMiner): هي منصة برمجيات علوم بيانات تم تطويرها من قبل الشركة التي تحمل هذا الاسم، وتحتوي على مجموعة من الوسائل للتنقيب في البيانات. إنه صعب الاستخدام، ولكن بمزيد من الممارسة، يمكن للمحلل الحصول بسرعة على سلسلة كاملة من معالجة البيانات.

ستاتا (Stata): إنها حزمة برمجيات ذات غاية إحصائية عامة، بحيث تمتد قدرات «الستاتا» لتشمل إدارة البيانات، والتحليل الإحصائي، والرسوم البيانية، وإعداد البرامج. ويتكون الاسم من كلمتي إحصاء وبيانات.

شعاع الدعم الآلي (Support Vector Machines): هي نماذج تعليم مراقب ذات خوارزميات تعليم مرتبط، يحلل البيانات لاستخدامها في التصنيف وتحليل الانحدار. وقد استخدمت هذه النماذج في شتى العلوم مثل علوم الأحياء لتصنيف البروتينات.

صلاحية متبادلة (Cross-Validation): هي تقنية تستعمل في تقييم كيفية تعميم نتائج تحليل إحصائي على مجموعة بيانات مستقلة.

طرق «بايز» الساذج (Naïve Bayes Methods): هي مجموعة من خوارزميات التعليم المراقب، القائمة على تطبيق نظرية بايز، في علاقتها بالافتراض الساذج للاستقلال بين كُلّ زوج من السمات على حدة. كما تعد طرقاً إحصائية للتصنيف.

غامب برو (JMP Pro): هو نسخة تحليلية متقدمة من «الغامب» الذي يمكننا من استخدام البيانات التي بحوزتنا لتوقع المستقبل بشكل أفضل والتخطيط له. و«الغامب برو» برمجية، يقدم كُلّ البيانات المتفوقة بشكل واضح.

فرضية صفرية أو (عدم) (Null Hypothesis): تمثل الفرضية الصفرية بـ H_0، وهي عادة فرضية تقوم بمعاينة ترصدات تنشأ صدفة. وهي تقوم على فكرة عدم وجود أي علاقة بين ظاهرتين تم قياسهما، أو أي ترابط بين مجموعات.

اللاسو (Least Absolute Shrinkage and Selection Operator) (LASSO): في علم الإحصاء، يعد اللاسو طريقة تحليل انحدار، تنجز عمليتي انتقاء المتغير والتضبيط بغية تحسين دقة التنبؤ وتفسير النموذج الإحصائي الذي تنتجه.

متغير (Variable): قد يكون المتغير شيئاً أو حدثاً، أو فكرة، أو شعور، أو فترة، أو أي فئة تحاول قياسها.

متغيرات تابعة (Dependent Variables): المتغير التابع هو ما يتم قياسه في التجربة، وهو الذي يتأثر خلال هذه التجربة. ويدعى تابع لأن وجوده «يتوقف» على وجود متغير مستقل؛ ومن ثم، لا يمكن تصور متغير تابع من دون متغير مستقل. إذا كنت مثلاً مهتماً بمقدار تأثير الضغط في معدل ضربات القلب لدى الإنسان، فسيكون متغيرك المستقل هو الضغط، ومتغيرك التابع هو معدل دقات القلب. ويمكنك بشكل مباشر معالجة مستويات الضغط لدى المبحوثين، وقياس كيفية تغيير مستويات الضغط، معدل دقات القلب.

متغيرات مستقلة (Independent Variables): هو متغير قائم بذاته ولا تغيره متغيرات أخرى. قد يكون العمر مثلاً متغيراً مستقلاً، ذلك بأن عوامل أخرى من قبيل

«نوع الأكل الذي يتناوله» صاحب هذا العمر، وكم من مرة يتردد على المدرسة، وكم من ساعة يشاهد فيها التلفاز، هي أمور لا تغير العمر.

متغيرات مستمرة (Continuous Variables): إذا تمكن متغير ما من أخذ أي قيمة بين قيمته القصوى وقيمته الدنيا، صار متغيراً مستمراً، وإذا أخفق، عُدّ متغيراً منفصلاً.

متغيرات وهمية أو صورية (Dummy Variables): هو متغير رقمي يستخدم في تحليل الانحدار لتمثيل مجموعات فرعية للعينة في البحث. وفي تصميم بحث ما، غالباً ما يستخدم متغير وهمي في التمييز بين مجموعات العلاج المختلفة.

مصفوفة ارتباك (Confusion Matrix): وتحتوي على معلومات حول التصنيفات الحقيقية والمتنبأة، التي تتم بواسطة نظام تصنيف ما. وإن أداء هذه النظم، تقيم عادة من خلال استعمال البيانات في المصفوفة. وتستمد مصفوفة الارتباك قوتها انطلاقاً من تحديدها لطبيعة تصنيف الأخطاء وكمياتها.

مصفوفة ترابطية أو علائقية (Correlation Matrix): تشير إلى جدول يعرض معاملات ترابطية بين مجموعات من المتغيرات. إنها تفحص طبيعة التبعية بين متغيرات متعددة في الوقت نفسه.

معامل تضخم التباين (VIF)(Variance Inflation Factor): عوامل تقيس مقدار تباين تضخم معاملات الانحدار المقدرة مقارنة بالحالة التي تكون فيها متغيرات المتنبئ غير مترابطة خطياً.

معاملات الانحدار (Regression Coefficients): معامل الانحدار في الإحصائيات، هو «a» الثابتة في معادلة الانحدار التي تخبرنا عن تغيير قيمة المتغير التابع الذي يوافق تغير الوحدة في المتغير المستقل.

معيار أكايكي للمعلومة (AIC) (Akaike Information Criterion): إنه معيار يستخدم في العديد من المجالات العلمية، لمقارنة جودة مجموعة من النماذج الإحصائية المتنافسة، وانتقاء الأنسب منها.

معيار بايز للمعلومة (Bayes information Criterion) (BIC): هو معيار يستخدم في انتقاء نموذج ما من بين مجموعة نماذج محدودة، بحيث ينتقى أساساً النموذج الذي لديه أقل نسبة من معيار بايز للمعلومة، وهو وثيق الصلة بمعيار أكايكي للمعلومة.

منحنى خاصية التشغيل المتلقي (Receiver Operating Characteristic) (ROC): هو رسم بياني، يوضح أداء نظام تصنيف ثنائي كلما تباينت عتبة التمييز. وهي طريقة تقارن الاختبارات التشخيصية. كما أنه رسم بياني يمثل معدل الإيجابي الصادق مقابل المعدل الإيجابي الكاذب.

نماذج الشبكات العصبية (Neural Network Models): الشبكة العصبية نموذج بيانات حاسوبية قوية، قادرة على ضبط وتمثيل العلاقات المدخلة والمخرجة المعقدة. وكان الدافع من وراء تطوير هذه الشبكة، هو تشكيل نظام اصطناعي يمكن أن يؤدي مهام ذكية شبيهة بذكاء عقل الإنسان، الذي يكتسب المعلومة ويخزنها.

نمذجة جزئية المربعات الكامنة الصغرى (Partial Least Squares Latent Modeling): هي طريقة إحصائية تستخدم في تشكيل نماذج تنبؤية عندما تكون العوامل متعددة وخطية مشتركة بشكل كبير، كما تستخدم لإيجاد علاقات أساسية بين مصفوفتين X و Y.

ثبت المصطلحات

Classification and Regression Trees (CART)	أشجار الانحدار والتصنيف
Decision Trees	أشجار القرار
k-Nearest Neigbours	k-أقرب الجيران
Stepwise Regression	انحدار تدريجي
Latent Class Regression	انحدار الطبقة الكامنة
Logistic Regression	انحدار لوجيستي
Multiple Regression	انحدار متعدد
Ordinary Least Squares Regression	انحدار المربعات الصغرى العادية
Data Dredging	تجريف البيانات
Computer Clusters	تجميعات الحاسوب
Principal Component Analysis (PCA)	تحليل المكوّن الرئيسي
Independent Component Analysis (ICA)	تحليل المكوّن المستقل
Bonferroni Correction	تصحيح بونفيروني
Machine Learning	تعلم آلي
Trees Partition	تقسيم الأشجار
Recursive Partitioning	تقسيم عَودي
JMP Pro	غامب برو
Statistical Package for the Social Sciences (SPSS)	حزمة إحصائية للعلوم الاجتماعية

Data Mining (DM)	تنقيب في البيانات
Kernel Trick	حيلة أو خدعة النواة
CHi-Squared Automatic Interaction Detector (CHAID)	مربع كاي للكشف عن التفاعل التلقائي
RapidMiner	«رابدماينر» أو منقب سريع
Stata	ستاتا
Support Vector Machines	شعاع الدعم الآلي
Cross-Validation	صلاحية متبادلة
Naïve Bayes Methods	طرق بايز الساذج
Null Hypothesis	فرضية صفرية أو (عدم)
Least Absolute Shrinkage and Selection Operator (LASSO)	لاسو
Variable	متغير
Dependent Variables	متغيرات تابعة
Independent Variables	متغيرات مستقلة
Continuous Variables	متغيرات مستمرة
Dummy Variables	متغيرات وهمية أو صورية
Confusion Matrix	مصفوفة ارتباك
Correlation Matrix	مصفوفة ترابطية أو علائقية
Variance-Inflation-Factor (VIF)	معامل تضخم التباين
Regression Coefficients	معاملات الانحدار
Akaike Information Criterion (AIC)	معيار أكايكي للمعلومة
Bayes Information Criterion (BIC)	معيار بايز للمعلومة
Receiver-Operating-Characteristic-(ROC)	منحنى خاصية التشغيل المتلقي
Neural Network Models	نماذج الشبكات العصبية
Partial Least Squares Latent Modeling	نمذجة جزئية المربعات الكامنة الصغرى

المراجع

Abbott, Andrew. 2001. Time Matters: On Theory and Method. Chicago, IL: University of Chicago Press.

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. 1993. «Mining Association Rules between Sets of Items in Large Databases.» Association for Computing Machinery (AMC) SIGMOD Record 22 (2):207-16.

Aiken, Leona S., and Stephen G. West 1991. Multiple Regression: Testing and Interpreting Interactions. Newbury Park, CA: Sage.

Albnan, Naomi S. 1992. «An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression.» American Statistician 46 (3):175-85.

Armstrong, J. Scott. «Significance Tests Harm Progress in Forecasting.» International Journal of Forecasting 23 (2):321-27.

Attewell, Paul, Scott Heil, and Liza Reisel. 2011. «Competing Explanations of Undergraduate Noncompletion.» American Educational Research Journal 48 (3): 536-59.

Benjamini, Yoav. 2010. «Simultaneous and Selective Inference:

Current Successes and Future Challenges.» Biometrical Journal 52 (6):708-21.

Berk, Richard A. 2006. «An Introduction to Ensemble Methods for Data Analysis.» Sociological Methods and Research 34 (3):263-95.

Berry, William D. 1993. Understanding Regression Assumptions. Newbury Park, CA: Sage.

Blyth, Colin R. 1972. «On Simpson's Paradox and the Sure-Thing Principle.» Journal of the American Statistical Association 67 (338): 364-66.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. «A Training Algorithm for Optimal Margin Classifiers.» In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 144-52. ACM.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. «Understanding Interaction Models: Improving Empirical Analyses.» Political Analysis 14(1):63-82.

Breiman, Leo. «Statistical Modeling: The Two Cultures.» 2001. Statistical Science 16(3): 199-231.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. CART: Classification and Regression Trees. Belmont, CA: Wadsworth.

Canty, Angelo, and Brian Ripley. 2012. boot: Bootstrap R (S-Plus) Functions, Version 1.2-42 (R package). http://cran.r-project.org/package=boot.

Coleman, James S., Thomas F. Pettigrew, William H. Sewell, and Thomas W. Pullum. 1973. «Inequality: A Reassessment of the Effect of Family and Schooling in America.» American Journal of Sociology 78 (6):1523-44.

Collins, Linda M., and Stephanie T. Lanza. 2010. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. Hoboken, NJ: John Wiley & Sons.

Comon, Pierre. 1994. «Indepe» Signal Processing 36(3): 287-314.

Cortes, Corinna, and Vladimir Vapnik. 1995. «Support-Vector Networks.» Machine Learning 20(3): 273-97.

Cover, Thomas, and Peter Hart 1967. «Nearest Neighbor Pattern Classification.» IEEE Transactions on Information Theory 13(1): 21-27.

Cui, Dapeng, and David Curry. 2005. «Prediction in Marketing Using the Support Vector Machine.» Marketing Science 24(4): 595-615.

Daniels, Ben. 2012. CROSSFOLD: Stata Module to Perfonn K-fold Cross-Validation (usergenerated Stata program). http://ideas.repec.org/c/boc/bocode/s457426.html.

Davison, Anthony Christopher, and David Hinckley. 1997. Bootstrap Methods and their Application. Cambridge: Cambridge University Press.

Deb, Partha. 2012. FMM: Stata Module to Estimate Finite Mixture Models (user-generated Stata program). http://ideas.repec.org/c/boc/bocode/s456895.html.

Dudani, Sahibsingh A. 1976. «The Distance-Weighted K-Nearest-Neighbor Rule.» IEEE Transactions on Systems, Man and Cybernetics 4:325-27.

Dunteman, George H. 1989. Principal Components Analysis. Newbury Park, CA: Sage.

Dupuis, Debbie J., and Maria-Pia Victoria-Feser. 2013. «Robust VIF Regression with Application to Variable Selection in Large Data, Sets.» Annals of Applied Statistics 7(I): 319-41.

Efron, Bradley. 1979. «Bootstrap Methods: Another Look at the Jackknife.» Annals of Statistics 7(1):1-26.

Efron, Bradley, and Gail Gong. 1983. «A Leisurely Look at the Bootstrap, the Jacklnife, and Cross-Validation.» American Statistician 37(1):36-48.

Elwert, Felix, and Christopher Winship. 2010. «Effect Heterogeneity and Bias in Main-Effects Only Regression Models.» In Heuristics, Probability and Causality: A Tribute to Judea Pearl, edited by Felix Elwert and Christopher Winship, 327-36. London: College Publications.

Esping-Andersen, Gosta. 1990. The Three Worlds of Welfare Capitalism. Cambridge: Polity.

Fawcett, Tom. 2006. «An Introduction to ROC Analysis.» Pattern Recognition Letters 27(8): 861-74·

Foster, Dean P., and Robert A. Stine. 2004. «Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy.» Journal of the American Statistical Society 99(461): 303-313.

Foster, Dean P., and Robert A. Stine. 2008. «α.-investing: a Procedure for Sequential Control of Expected False Discoveries.» Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(2):429-44.

Freedman, David A. 2006. «On the So-Called «Huber Sandwich Estimator» and «Robust Standard Errors.»»» American Statistician 60 (4):299-302.

Gavrishchaka, Valeriy V., and Supriya Banerjee. 2006. «Support Vector Machine as an Efficient Framework for Stock Market Volatility Forecasting.» Computational Management Science 3(2):147-60.

Goeman, Jelle J. 2010. «Lr Penalized Estimation in the Cox Proportional Hazards Model.» Biometrical Journal 52 (1): 70-84.

Goeman, Jelle J., Rosa Meijer, and Nimisha Chaturvedi. 2012. L1 and L2 Penalized Regression Methods (R package). http://cran.r-project.org/web/packages/penalized/vignettes/ penalized.pdf

Halko, Nathan, Per-Gunnar Martinsson. and Joel A. Tropp. 2011. «Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions.» SIAM Review 53(2):217-88.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. Data Mining: Concepts and Techniques. 3rd ed. New York: Elsevier.

Haralick, Robert, and Rave Harpaz. 2007. «Linear Manifold Clustering in High Dimensional Spaces by Stochastic Search.» Pattern Recognition 40(10):2672-84.

Hastie, Trevor, and Robert Tibshirani. 1996. «Discriminant Adaptive Nearest Neighbor Classification.» IEEE Transactions on Pattern Analysis and Machine Intelligence 18(6): 607-16.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, and James Franklin. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

Hsu, Jason C. 1996. Multiple Comparisons: Theory and Methods. New York: Chapman & Hall.

Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja. 2001. Independent Component Analysis. New York: John Wiley & Sons.

Hyvärinen, Aapo, and Erkki Oja. 2000. «Independent Component Analysis: Algorithms and Applications.» Neural Networks 13(4):4II-30.

Ioannidis, John PA. 2005. «Why Most Published Research Findings are False.» PLoS Medicine2(8):eI24.

Jaccard, James, and Robert Turrisi. 2003. Interaction Effects in Multiple Regression. 2nd ed. Thousand Oaks, CA: Sage.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns, and Stephan Michelson.

1972. Inequality: A Reassessment of the Effects of Family and Schooling in America. New York: Basic Books.

Kostaki, Anastasia, Javier M. Moguerza, Alberto Olivares, and Stelios Psarakis. 2012. «Support Vector Machines as Tools for Mortality Graduation.» Canadian Studies in Population 38(3-4):37-5 8.

Kuhn, Max, and Kjell Johnson. 2013. Applied Predictive Modeling. New York: Springer.

Kuhn, Thomas S. 1962. The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. New York: Wiley.

Larson, Selmer C. 1931. «The Shrinkage of the Coefficient of Multiple Correlation.» Journal of Educational Psychology 22(1):45.

Larzelere, Robert E., and Stanley A. Mulaik. 1997. «Single-Sample Tests for Many Correlations.» Psychological Bulletin 84(3):557.

Laub, John H., Daniel S. Nagin, and Robert J. Sampson. 1998. «Trajectories of Change in Criminal Offending: Good Marriages and the Desistance Process.» American Sociological Review 63(2):225-38.

Lazarsfeld, Paul Felix, and Neil W. Henry. 1968. Latent Structure Analysis. New York: Houghton Mifflin.

Lewis, David D. 1998. «Naive (Bayes) at Forty: The Independence Assumption in Infonnation Retrieval.» In Machine Learning: ECML-98, 4-15. Berlin: Springer.

Lin, Dongyu. 2011. VIF Regression: A Fast Regression Algorithm for Large Data (R package). http://cran.r-projectorg/web/packages/VIF/VIF.pdf

Lin, Dongyu, Dean P. Foster, and Lyle H. Ungar. 2011. «VIF

Regression: A Fast Regression Algorithm for Large Data.» Journal of the American Statistical Association 106(493):232-47.

Linzer, Drew A., and Jeffrey B. Lewis. 20ll. «poLCA: An R Package for Polytomous Variable Latent Class Analysis.» Journal of Statistical Software 42(10) :1-29.

Marchini, Jonathan L., Christopher Heaton, and Brian D. Ripley. 2012. FastICA Algorithms to Perform ICA and Projection Pursuit (R package). http://cran.r-projectorg/web/packages/ fastICAjfastICA.pdf

Martinsson, Per-Gunnar, Vladimir Rokhlin, and Mark Tygert 2011. «A Randomized Algorithm for the Decomposition of Matrices.» Applied and Computational Harmonic Analysis 30(1): 47-68..

McKinsey Global Institute. 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. http://www.mckinsey.com/ insights/business-technology/big-data-the -next- frontier-for-innovation.

Melamed, David, Ronald L. Breiger, and Eric Schoon. 2013. «The Duality of Clusters and Statistical Interactions.» Sociological Methods and Research 42(1):41-59.

Miller, Alan. 2002. Subset Selection in Regression. 2nd Edition. Boca Raton, FL: Chapman and Hall/ CRC.

Mooney, Christopher Z., and Robert D. Duval. 1993. Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, CA: Sage.

Morrison, Denton E., and Ramon E. Henkel. 1970. The Significance Test Controversy: A Reader. New Brunswick, NJ: Transaction.

Murphy, Kevin. 2012. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press.

Nickerson, Raymond S. 2000. «Null Hypothesis Significance Testing:

A Review of an Old and Continuing Controversy.» Psychological Methods 5(2): 241.

Nisbet, Robert, John Elder IV, and Gary Miner. 2009. Handbook of Statistical Analysis and Data Mining Applications. New York: Academic.

North, Matthew. 2012. Data Mining for the Masses. [United States:] Global Text Project.

Pearl, Judea. 2000. Causality: Models, Reasoning, and Inference. New York: Cambridge University Press.

Qian, Bo, and Khaled Rasheed. 2010. «Foreign Exchange Market Prediction with Multiple Classifiers.» Journal of Forecasting 29(3):271-84.

Ragin, Charles C. 2008. Redesigning Social Inquiry: Fuzzy Sets and Beyond. Chicago: University of Chicago Press.

Ridgeway, Greg. 1999. «The State of Boosting.» Computing Science and Statistics 31:172-81.

Rish, Irina. 2001. «An Empirical Study of the Naive Bayes Classifier.» In IJCAI2001 Workshop on Empirical Methods in Artificial Intelligence, 41-46.

Rubin, Donald B. 1978. «Bayesian Inference for Causal Effects: The Role of Randomization.» Annals of Statistics 6(1):34-58.

Ruger, T. W., Kim, P. T., Martin, A. D., & Quinn, K. M. 2004. «The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking.» Columbia Law Review 104(4):1150-1210.

Saville, Dave J. 1990. «Multiple Comparison Procedures: The Practical Solution.» American Statistician 44(2):174-80.

Schonlau, Matthias. 2005. «Boosted Regression (Boosting): An Introductory Tutorial and a Stata Plugin.» Stata Journal 5(3):330.

Shaffer, Juliet Popper. 1995. «Multiple Hypothesis Testing.» Annual Review of Psychology 46 (1):561-84.

Silver, Nate. 2012. The Signal and the Noise: Why So Many Predictions Fail-But Some Don't. New York: Penguin.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining. Boston: Addison-Wesley.

Taleb, Nassim N. 2005. Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets. New York: Random House.

------.2007. The Black Swan: The Impact of the Highly Improbable. New York: Random House.

Thomas, Scott L., and Ronald H. Heck. 2001. «Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated with Complex Sampling Designs.» Research in Higher Education 42(5):SI7-40.

Tibshirani, Robert. «Regression Shrinkage and Selection via the LASSO.» 1996. Journal of the Royal Statistical Society, Series B (Methodological) 58(1):267-88.

Tukey, John W. 1991. «The Philosophy of Multiple Comparisons.» Statistical Science 6(1):100-16.

Vempala, Santosh S. 2004. The Random Projection Method. Providence, RI: American Mathematical Society.

Weerts, David J., and Justin M. Ronca. 2009. «Using Classification Trees to Predict Alumni Giving for Higher Education.» Education Economics 17(I):95-122.

Williams, Graham. 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. New York: Springer.

Williams, Richard. 2010. Estimating Heterogeneous Choice Models

with oglm. Department of Sociology, University of Notre Dame. http://www3.nd.edu/-rwilliam/oglm/oglnLStata .pdf.

Willis, Paul. 1977. Learning to Labour: How Working-Class Kids Get Working-Class Jobs. Farborough: Saxon House.

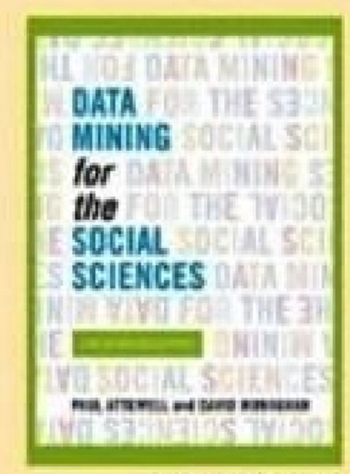
Witten, Ian H., Eibe Frank, and Mark Hall. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. New York: Morgan Kaufmann.

Xu, Lei, Adam Krzyzak, and Ching Y. Suen. 1992. «Methods of Combining Multiple Classifiers and their Applications to Handwriting Recognition.» IEEE Transactions on Systems, Man and Cybernetics 22(3): 418-35.

Zhang, Harry. 2004. «The Optimality of Naive Bayes.» In Proceedings of the Seventeenth International FLAIRS Conference, 562-67.

الفهرس

مدخل إلى التنقيب في بيانات العلوم الاجتماعية



- أسوار المعرفة العلمية
- ثقافة علمية معاصرة
- فلسفة
- علوم إنسانية واجتماعية
- تقنيات وعلوم تطبيقية
- آداب وفنون
- لسانيات ومعاجم

الحفر في البيانات، أو ما يطلق عليه أحياناً اسم استكشاف البيانات أو المعرفة، عملية من عمليات تحليل البيانات، وتلخيصها ضمن معلومات مفيدة، قد تستخدم مثلاً في الرفع من الدخل، أو تخفيض التكاليف أو هما معاً. وإن برمجيات الحفر في البيانات هي إحدى الوسائل التحليلية العديدة المسخرة في عملية الحفر في البيانات؛ فهي تمكن المستخدمين من تحليل البيانات انطلاقاً من أبعاد وزوايا مختلفة، وتصنيفها، وتلخيص العلائق المرصودة. ومن الناحية التقنية، يعد الحفر في البيانات، عملية تحدد التعالقات أو الأنماط الموجودة بين عشرات الحقول في قواعد البيانات العلائقية الضخمة. ومع ذلك يظل هذا التحليل محدوداً بالمقارنة مع ما وصلت إليه الابتكارات المستمرة في مجال المعالجة الحاسوبية، وتخزين القرص، والبرمجيات الإحصائية التي رفعت من دقة تحليل البيانات على نحو لافت للنظر. وقد تكون البيانات وقائع، أو أعداداً، أو نصاً يمكن أن تخضع إلى المعالجة الحاسوبية، كما أن التقدم الذي تم تحقيقه في مجال برمجيات الحاسوب، مكّنت المنظمات والشركات، وغيرها، من دمج قواعد بياناتها في مستودع البيانات؛ إذ تدار داخله البيانات بشكل منظم وتسترجع متى شاء المحلل ذلك. ومن بين هذه البرمجيات التحليلية، نذكر البرمجيات الإحصائية، وبرمجيات التعلم الآلي، وبرمجيات الشبكات العصبية، بحيث تسعى كلها إلى البحث في «الأصناف»، و«التجميعات»، و«الترابطات»، و«الأنماط التسلسلية».

المؤلف: بول أتيويل: أستاذ متميز في علم الاجتماع في مركز الدراسات العليا بجامعة مدينة نيويورك، حيث يعطي دروساً حول تحليل البيانات.

دايفيد ب. موناغان: مرشح دكتوراه في علم الاجتماع في مركز الدراسات العليا في جامعة مدينة نيويورك، وأعطى دروساً حول طرق البحث الكمي والديمغرافيا والتعليم.

المترجم: عبد النور خراقي: أستاذ اللغويات، والمدير المسؤول عن تكوين الماجستير في «التواصل، والثقافة، والترجمة» بقسم اللغة الإنجليزية، بكلية الآداب والعلوم الإنسانية، وجدة، المغرب، والمسؤول عن التكوينات المستمرة بالإنجليزية «بمركز اللغات والتواصل» التابع لجامعة محمد الأول بوجدة، المغرب.



المنظمة العربية للترجمة



الثمن: 23 دولاراً
أو ما يعادلها